

# Challenges in AI-driven multi-omics data analysis for Oncology: Addressing dimensionality, sparsity, transparency and ethical considerations

Maryem Ouhmouk<sup>a,\*</sup>, Shakuntala Baichoo<sup>b</sup>, Mounia Abik<sup>a,c</sup>

<sup>a</sup> Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes (ENSIAS), Mohammed V University in Rabat, Rabat, Morocco

<sup>b</sup> Department of Digital Technologies, University of Mauritius, Reduit, Mauritius

<sup>c</sup> The Royal Institute for Training of Youth and Sports Executives (IRFC / JS), Rabat, Morocco

## ARTICLE INFO

### Keywords:

Multi-omics  
genomics  
Data integration  
AI  
Deep learning  
Ethics  
Privacy  
Fairness  
Cancer  
Healthcare systems

## ABSTRACT

Artificial intelligence, particularly deep learning, is becoming increasingly prominent in multi-omics research, especially since traditional statistical models struggle to handle the complexity and high dimensionality of such data. By effectively combining different types of omics data, AI techniques can unveil hidden connections, detect biomarkers, and improve disease prediction through the integration of multi-omics layers and modalities, which can lead to significant advancements in precision medicine. In this review, we gathered published methods of deep learning-based multi-omics integration specialized in oncology since 2020. We concentrated exclusively on studies utilizing cancer omics data mainly sourced from The Cancer Genome Atlas (TCGA) database. As a result, we identified 32 articles that generally fulfilled the criteria. We studied their techniques and their ability to handle challenges in analyzing multi-omics data, particularly regarding missing data, dimensionality, and processing workflows. We also discuss how well these methods consider explainability, interpretability, and ethical aspects in developing solutions that treat private medical and sensitive information.

From the 32 studies, we can divide deep learning-based multi-omics integration methods into two types: non-generative and generative models. Non-generative approaches, such as feedforward neural networks (FFNs), graph convolutional networks (GCNs), and autoencoders, are designed to extract features and perform classification directly. On the other hand, generative methods such as variational autoencoders (VAEs), generative adversarial networks (GANs), and generative pretrained transformers (GPTs) focus on creating adaptable representations that can be shared across multiple modalities. These methods have advanced the handling of missing data and dimensionality, outperforming traditional approaches. However, most reviewed models remain at the proof-of-concept stage, with limited clinical validation or real-world deployment.

## 1. Introduction

Multi-omics analysis integrates information from genomics, transcriptomics, proteomics, and metabolomics to provide a more comprehensive picture of biological functions, disease pathways, and possible therapeutic approaches. Each molecular layer offers distinct perspectives, but together they enable a more detailed understanding of cellular mechanisms, signaling cascades, and expression trends. Still, merging these layers poses significant challenges. Variations in data formats, measurement scales, and reliability across different platforms demand sophisticated techniques to extract valuable insights [1].

Artificial intelligence (AI), including machine learning (ML) and deep neural networks (DL), is reshaping the landscape of healthcare by

enhancing the ways diseases are diagnosed, forecasted, and managed [2]. These computational methods are particularly effective in dissecting large, multi-dimensional omics datasets, often uncovering associations and patterns that conventional statistical tools overlook. The result is improved diagnostic accuracy, novel target identification, and greater potential for individualized treatment strategies [3]. Despite these advancements, difficulties persist such as harmonizing heterogeneous data sources, the need for large annotated cohorts, and the challenge of maintaining interpretability. Addressing these problems is essential for responsible and effective adoption of AI in medical research [4].

AI-driven approaches to multi-omics integration offer considerable promise for advancing personalized medicine. By synthesizing multiple biological data types, AI systems can discern intricate relationships,

\* Corresponding author.

E-mail address: [maryem\\_ouhmouk@um5.ac.ma](mailto:maryem_ouhmouk@um5.ac.ma) (M. Ouhmouk).

<https://doi.org/10.1016/j.imu.2025.101679>

Received 16 May 2025; Received in revised form 20 July 2025; Accepted 1 August 2025

Available online 4 August 2025

2352-9148/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

anticipate disease patterns, and pinpoint biomarkers with greater efficiency than many traditional workflows, especially when handling complex, high-dimensional data. However, performance often depends on data quality, the difficulty of the analytic task, and the rigor of validation procedures. Progress in this domain also introduces ethical considerations, including patient data privacy, information security, and potential algorithmic biases. The development of transparent guidelines and regulatory standards is vital to ensure ethical use and safeguard patient interests. Promoting model interpretability through explainable AI methods and incorporating up-to-date clinical inputs can further encourage confidence and wider implementation in healthcare settings [5,6].

Nevertheless, as emphasized in this review, most AI-based multi-omics platforms are still largely confined to proof-of-concept stages, with minimal uptake in actual clinical practice. Despite encouraging results in areas such as subtype discrimination or survival estimation, these approaches frequently lack robust external validation, real-world clinical trials, or regulatory approval.

In summary, the integration of AI with multi-omics data analysis provides a powerful avenue for improving our understanding and treatment of diseases. By tackling technical limitations, improving model transparency, and proactively managing ethical risks, the field can unlock the full benefits of these technologies. This will ultimately pave the way toward more targeted, personalized, and accessible healthcare, supporting the ongoing progress of precision medicine.

## 2. Multi-omics data foundations

Traditional statistical models are often used to analyze and interpret large datasets. However, over the last decade, AI has emerged as a powerful tool in a variety of fields, revolutionizing data analysis and decision-making processes. This shift is driven by the increasing complexity of data, which comes from traditional structured forms to highly complicated and multi-dimensional datasets. The growing demand for deeper understanding of biological processes has accelerated the need for more advanced approaches in genomics data analysis [7].

Fig. 1 illustrates the key components of a successful AI-driven healthcare system, highlighting their relationship in achieving optimal performance. Starting with security, which is critical for protecting

patient data that supports data integrity and ensures the reliability and accuracy of healthcare data. In addition, ethics and privacy are closely related because of their ability to advocate for responsible AI, whereas transparency and explainability increase trust by making AI workflows understandable. Furthermore, bias reduction and fairness are critical for ensuring that AI models produce equitable healthcare outcomes. All of these factors combined with clinical integration to ensure that AI solutions are seamlessly incorporated into daily medical practice, resulting in a more effective and trustworthy AI-powered healthcare system [7–10].

Different types of multi-omics data offer a comprehensive view of biological systems in AI-driven healthcare [11]. Genomics analyzes DNA variation, transcriptomics profiles RNA expression, proteomics quantifies proteins, metabolomics measures metabolites, epigenomics examines regulatory marks, and microbiome profiling investigates host–microbe interactions. Integrated with AI, these layers enable more precise and personalized insights (see Fig. 2).

Since AI was formally introduced, it expanded into a powerful domain that includes techniques like ML and DL. These methods enable computers to not only process large volumes of data but also learn from it. ML, in particular, is focused on developing algorithms and models that allow systems to perform tasks without the need for explicit instructions. The selection of the appropriate algorithm is key, and these can be broadly classified into three categories: supervised, semi-supervised, and unsupervised learning [12]. Meanwhile, deep learning is a subset of machine learning that uses layered neural networks to model complex patterns in large datasets. It mainly excels in tasks that involve high-dimensional data, such as image and speech recognition, due to its ability to automatically extract hierarchical features without manual feature engineering [4,13].

Several key databases are used in ML and DL based multi-omics studies to provide diverse biological data such as genomic, transcriptomic, proteomic, epigenomic, and metabolomic information. These databases are very informative for building predictive models and improving our understanding of complex diseases like cancer (see Table 1).

The Cancer Genome Atlas (TCGA) [14] is a widely used resource in multi-omics integration, offering comprehensive molecular data from over 30 cancer types, including DNA mutations, RNA expressions,

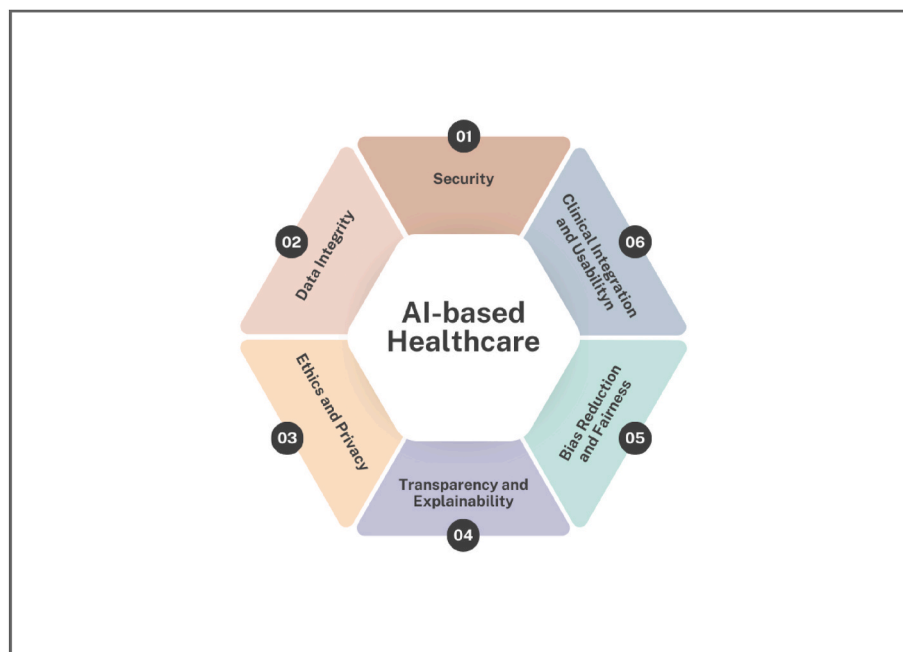


Fig. 1. Aspects of successful AI-based healthcare.

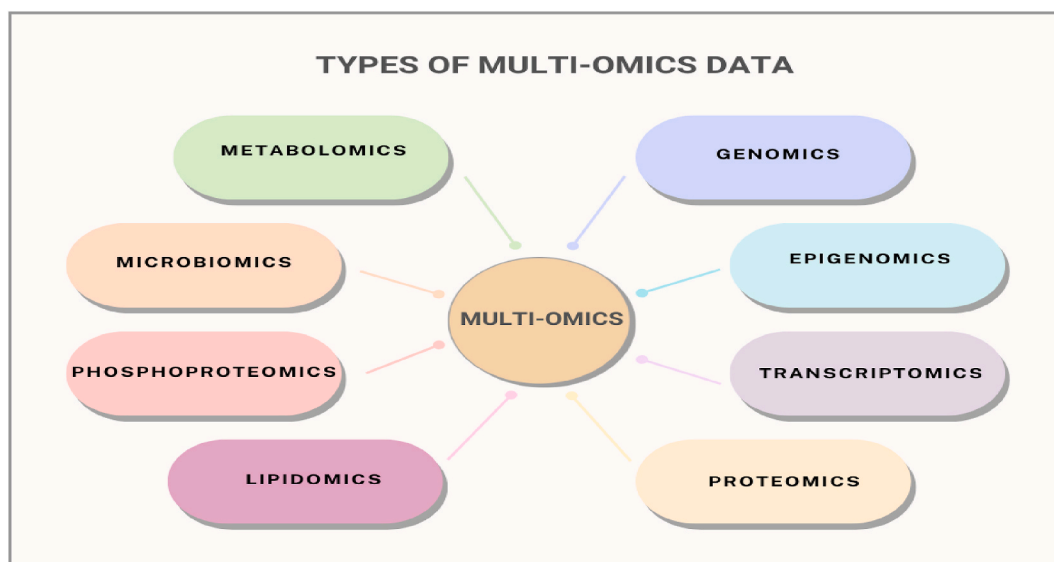


Fig. 2. Different types of multi-omics data.

Table 1  
Databases for multi-omics data.

Database	Source	No. of Samples	Type of Data
TCGA	Cancer tissue samples across 33 cancer types	>20,000 samples	Genomic (mutations, CNVs), transcriptomic (mRNA, miRNA), epigenomic (methylation), proteomic.
METABRIC	Breast cancer tissue samples	>2000 breast cancer samples	Genomic (mutations, CNVs), transcriptomic (mRNA expression), clinical data).
CCLE	Cancer cell lines from various tissues	>1400 cell lines	Genomic (mutations, CNVs), transcriptomic, drug response data
TARGET	Pediatric cancer tissue samples	>3000 pediatric cancer cases	Genomic (mutations, CNVs), transcriptomic, methylation, clinical data.
GEO	Public repository for gene expression studies	>2 million samples	Transcriptomic (mRNA, miRNA, non-coding RNA expression), epigenomic, genomic, proteomic.
GTEx	Normal human tissues from various donors	>17,000 samples across 54 tissues	Genomic (SNPs), transcriptomic (gene expression), eQTLs (expression quantitative trait loci).
The 1000 Genomes project	Genomic DNA from blood or other non-disease-specific tissues of healthy donors.	>2500 individuals from 26 populations	Genomic (mutations, CNVs), transcriptomic, methylation, clinical data.
The International Cancer Genome Consortium (ICGC)	tumor samples and matching normal tissues from cancer patients.	>25,000 cancer genomes	Genomic (mutations, CNVs), transcriptomic, methylation, clinical data.

methylation, and copy number variations (CNVs). Another famous database is the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [15], which contains genomic and transcriptomic data from over 2000 breast cancer patients, making it a key source for studies on cancer subtyping and prognosis using ML/DL approaches.

The Cancer Cell Line Encyclopedia (CCLE) [16] provides data on hundreds of cancer cell lines, including genomic alterations and drug response profiles. The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [17] focuses on pediatric cancers. Furthermore, Gene Expression Omnibus (GEO) [18], one of the largest public repositories, stores gene expression data across various conditions and species. The International Cancer Genome Consortium (ICGC) [19] and the 1000 Genomes Project [20] are both large-scale collaborative projects aimed at understanding the genomic changes in various types of cancer and collect genomic data from individuals representing diverse populations worldwide. Lastly, the Genotype-Tissue Expression Project (GTEx) [21] database links genetic variants to gene expression across different human tissues, providing insights into gene regulation and disease mechanisms.

By integrating data from these repositories, ML and DL models can learn complex patterns across multiple layers of biological information, supporting more accurate disease prediction and biomarker discovery. However, rigorous external validation remains essential to ensure their reliability and clinical applicability.

### 3. Challenges in multi-omics integration

In multi-omics research, data integration challenges make it challenging to understand how different datasets work together and affect results. These obstacles can limit a clear, unified view across diverse omics layers, stopping comprehensive conclusions. Developing robust integration platforms is essential to manage this complexity and unlock deeper biological understanding (see Fig. 3).

#### 3.1. High dimensionality and data sparsity

Omics data, which include high-throughput molecular information from genomics, transcriptomics, proteomics, and metabolomics, have greatly enhanced our understanding of biological processes and disease mechanisms, advancing precision medicine strategies. However, these datasets are often characterized by a limited number of samples relative

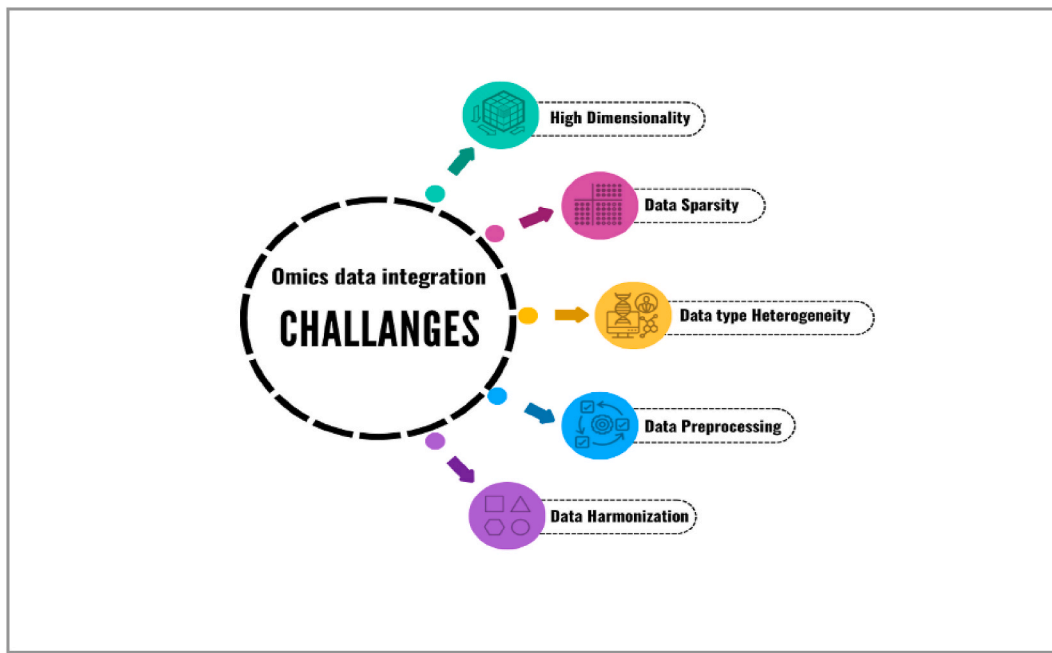


Fig. 3. Challenges of multi-omics data integration.

to the number of features, resulting in high dimensionality and data sparsity. This imbalance leads to several challenges: Distance-based metrics become unreliable, models tend to capture noise rather than meaningful biological patterns, and missing or undetected values further complicate analysis, often causing overfitting and unreliable predictions [22,23].

To address these obstacles, various methodological solutions are available. Techniques for reducing data dimensionality such as Principal Component Analysis (PCA) [24], t-distributed Stochastic Neighbor Embedding (t-SNE) [25], Uniform Manifold Approximation and Projection (UMAP) [26], TriMAP [27], LargeVis [28], and autoencoder networks [29], streamline complex datasets while retaining critical biological features. Penalization methods like L1 (Lasso) and L2 (Ridge) regularization help prevent model overfitting by controlling the flexibility of the predictive algorithms. Missing data can be managed using imputation strategies, including k-nearest neighbors (kNN) and matrix factorization techniques, thereby enhancing both the integrity of the dataset and the reliability of downstream analyses. Collectively, these approaches facilitate the robust integration and meaningful interpretation of multi-omics information.

### 3.2. Handling heterogeneous data types

Multi-Omics data consists of various biological data types, such as categorical data, continuous data, and ordinal data. These data types are generated by different technologies, each with unique formats, scales, and statistical distributions, making their integration a significant challenge [30]. One of the primary challenges lies in the different scales of the data. Omics data often vary in scale. For instance, gene expression levels may span a wide range, while others may simply be 0 or 1. Aligning these datasets onto a common scale can be difficult. Additionally, there are variable data structures to consider, as each omics layer follows distinct structural formats. For instance, protein-protein interactions may be represented as networks, whereas gene expression data are organized in a structured form. Moreover, the biological context of different omics data types can vary, making it challenging to combine them in a way that accurately reflects the underlying biological interaction [30].

To address these challenges, data transformation techniques, such as

standardization or normalization (for instance, z-score normalization or quantile normalization), can help harmonize different scales across omics layers [31,32]. Multi-layer networks is another approach, knowing that network-based methods, such as heterogeneous network integration, allow for the modeling of various biological interactions (including gene-gene or protein-protein (PPI)), addressing structural differences [33]. Lastly, matrix factorization techniques, including non-negative matrix factorization (NMF) and multi-view canonical correlation analysis (CCA), can identify shared patterns across different omics layers, which enables a more effective integration of heterogeneous data types [34,35]. More recently, Self-normalizing neural networks (SNNs) are more and more popular. They are particularly interesting because of their ability to enable high-level abstract representations without the need for batch normalization by using scaled exponential linear units (SELUs) to automatically maintain neuron activations at zero mean and unit variance [36].

### 3.3. Data preprocessing and harmonization

Data preprocessing and harmonization are critical steps in preparing multi-omics data for integration. This process involves cleaning the data, removing noise, and ensuring consistency across various datasets. Without proper preprocessing, the accuracy and quality of downstream analyses can be significantly compromised, leading to unreliable results.

One major challenge is batch effects, which arise when data is collected from different batches, labs, or platforms, which often leads to systematic differences caused by technical variation rather than biological factors which affect biological signals [37]. Data cleaning is another critical challenge, as datasets often include missing values, outliers, or noisy measurements that can bias results if not handled appropriately. Additionally, inconsistent annotations across datasets are very common as well [38]. Moreover, normalization and scaling across platforms pose significant challenges due to technology-specific biases inherent in different omics approaches [39].

To address batch effects, methods like ComBat (R function) or Remove Unwanted Variation (RUV) (R function) can help reduce technical variability, allowing biological signals to emerge more clearly [40–42]. For handling missing data and noise, advanced imputation techniques such as multiple imputations by chained equations (MICE)

[43] are mostly used, in addition to smoothing techniques that can reduce noise without losing important biological patterns. Ontology mapping, using standardized ontologies like Gene Ontology (GO) [44] or databases like UniProt [45] and Ensembl [46], can resolve inconsistent annotations, facilitating more accurate integration of cross-platform data. Finally, cross-omics normalization techniques, such as quantile normalization or variance-stabilizing transformation [37,47], can be applied to align datasets from different omics technologies onto a common scale, ensuring data from different platforms is compatible.

Addressing these challenges in multi-omics integration is critical for extracting meaningful insights from complex, multi-layered biological data. Techniques such as dimensionality reduction, network-based approaches, and robust data harmonization methods can mitigate some of these issues, but further advancements in both AI methodologies and omics technologies are needed to fully understand the potential of integrated multi-omics analysis. Table 2 lists some of the main of those challenges, providing some of the latest publications addressing them.

#### 4. AI methodologies for multi-omics

This review examines recent (2020–2024) developments in deep learning approaches for multi-omics data integration in cancer research, a period marked by rapid advances in sequencing technologies, computational frameworks, and open-source tools. We focused on this timeframe to ensure a current and targeted analysis, with particular attention to methods that show translational potential. To maintain consistency across studies, we prioritized those using data from The Cancer Genome Atlas (TCGA). Alongside methodological performance, we also considered ethical and practical aspects such as model transparency, fairness, and clinical applicability.

Our selection process involved a systematic search on PubMed and Google Scholar using keyword combinations like “multi-omics,” “deep learning,” “cancer,” and “TCGA” After removing around 50 duplicates, we screened titles and abstracts to retain studies centered on TCGA-derived multi-omics datasets. From an initial pool of 50 articles, we selected 32 based on methodological robustness and clarity of presentation (see Fig. 4).

While peer-reviewed publications were given priority, we also included preprints from platforms like bioRxiv and arXiv when they offered sufficient methodological transparency.

Integrating multi-omics data brings several challenges, including high dimensionality, missing values, heterogeneous modalities, and

batch effects. To tackle these, recent models have adopted deep learning strategies that focus on robust feature representation, modality fusion, and cross-modal generalization. We categorized these methods into three main groups, generative, non-generative, and hybrid, based on their architectural frameworks and integration logic. Each class addresses distinct aspects of the fusion problem. Fig. 5 illustrates the conceptual grouping, while Table 3 summarizes the characteristics, applications, and evaluation metrics of the 32 selected models.

This section explores how these model types, through techniques such as imputation, attention mechanisms, graph-based reasoning, and latent space modeling, are used to overcome the biological and computational complexities of multi-omics integration in oncology.

##### 4.1. Generative models

Generative models represent a class of machine learning frameworks designed to capture the underlying structure or probability distributions within data. In essence, these models enable systems to synthesize new data instances based on the patterns learned from training examples. Their application has expanded within multi-omics research, where they help to overcome issues like limited sample sizes, high feature dimensionality, and inconsistencies between data types. By extracting underlying, low-dimensional representations from complex omics datasets, generative models support more robust data integration and downstream analysis. Recent developments have focused on architectures such as **Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Generative Pretrained Transformers (GPTs).**

Variational Autoencoders (VAEs) are especially valued for their ability to generate concise and informative latent spaces from high-dimensional data. For instance, the MetaCancer pipeline [67] initiates preprocessing by discarding features with more than 25 % missing entries and removing samples lacking at least 75 % of features. It employs the R ‘impute’ function, using nearest neighbor averaging, to fill in gaps before feeding the data into a convolutional VAE. Likewise, OmiEmbed [69] utilizes mean imputation for missing molecular measurements and applies normalization before data embedding. In TMO-Net [70], gene expression matrices are filtered to remove genes with low variance (standard deviation < 1), while copy number variation (CNV) data retains segments with fewer than 5 % zero values, reducing background noise before integrating modalities. MoVAE [68] excludes features missing in over 20 % of samples, imputes remaining gaps with k-nearest neighbors, and uses z-score normalization to standardize across gene

**Table 2**  
Challenges of Multi-omics integration.

Challenge	Description	Solution	References
<b>Curse of Dimensionality</b>	High-dimensional datasets result in sparse data, making distance-based metrics less effective and reducing model reliability.	Dimensionality reduction techniques (PCA, t-SNE, Autoencoders) to reduce feature space while preserving key information.	[48,49]
<b>Overfitting</b>	A high number of features relative to the sample size increases the risk of overfitting, causing models to capture noise rather than biological patterns.	Regularization techniques (L1, L2) to penalize complexity and reduce overfitting.	[50,51]
<b>Sparsity of Data</b>	Missing or undetected measurements in multi-omics datasets lead to data sparsity, complicating the discovery of patterns or predictions.	Imputation methods (k-Nearest Neighbors, matrix completion algorithms) to manage missing data.	[38,52]
<b>Different Scales</b>	Different data types, such as binary genomic data and continuous gene expression data, complicate integration due to varying scales.	Data transformation techniques (z-score normalization, quantile normalization) to harmonize data scales.	[32,53]
<b>Variable Data Structures</b>	Diverse omics layers (networks for protein-protein interactions vs. tabular gene expression data) create challenges for integration.	Multi-layer networks and heterogeneous network integration to model various biological interactions in a unified framework.	[54–56]
<b>Biological Context</b>	Systematic differences across data from different labs, batches, or platforms can obscure true biological signals.	Batch effect correction methods (ComBat, RUV) to minimize technical variability and highlight biological signals.	[37,57]
<b>Data Cleaning</b>	Issues like missing values, outliers, and noisy measurements in real-world datasets can bias results if not addressed.	MICE and smoothing techniques to clean and refine the dataset for more reliable analysis.	[58,59]
<b>Inconsistent Annotations</b>	Variations in gene/protein/metabolite identifiers across platforms can lead to integration errors.	Ontology mapping using standardized resources (GO, UniProt, Ensembl) to resolve annotation inconsistencies.	[60,61]
<b>Normalization Across Platforms</b>	Different omics technologies introduce inherent biases, such as differences in sensitivity or dynamic range, requiring correction before integration.	Cross-omics normalization (quantile normalization, variance-stabilizing transformation) to align data for comparison.	[39,62]

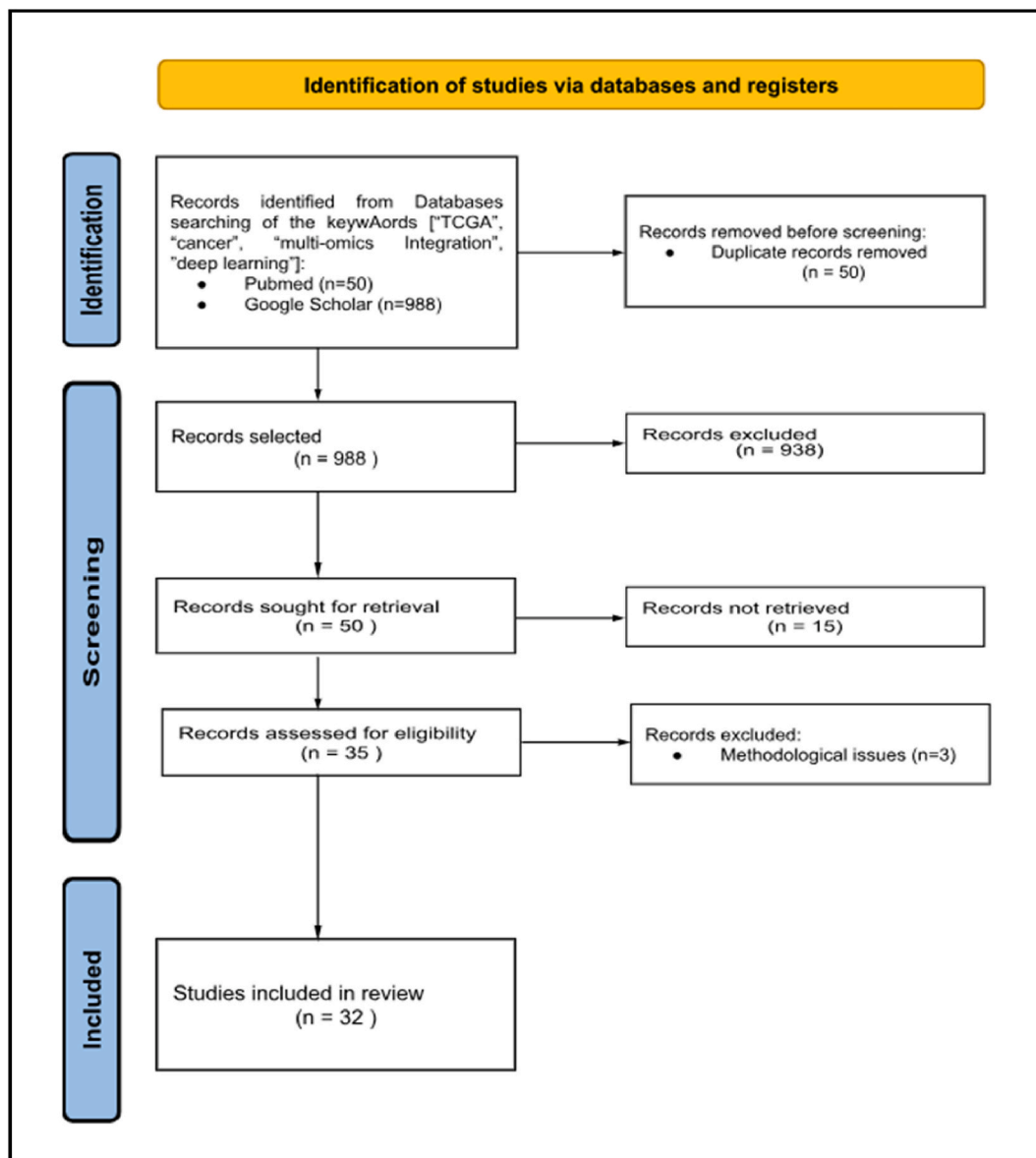


Fig. 4. PRISMA flow chart of screening process.

expression, methylation, and CNV profiles.

Generative Adversarial Networks (GANs) also play a key role in preprocessing through data augmentation and normalization. ctGAN [65] selects survival-associated genes via Cox proportional hazards (CoxPH) models and augments limited datasets by learning to replicate patterns observed in larger cohorts such as breast cancer (BRCA). Subtype-GAN [64] normalizes input distributions separately for mRNA, miRNA, copy number, and methylation data, integrating batch normalization layers to stabilize training and reduce mode collapse risks.

Generative Pretrained Transformers (GPTs) like mosGraphGPT [73] preprocess multi-omics data, for example in Alzheimer's disease research, by aligning samples, standardizing gene counts, and constructing knowledge graphs derived from KEGG pathways. PATH-GPTOMIC [72] refines bulk RNA-seq data using the scGPT model [94], which was initially trained on single-cell RNA-seq data. It leverages the pretrained backbone to generate a smooth latent space for bulk RNA-seq embeddings and introduces a smoothing module to address distribution shifts between single-cell and bulk transcriptomic data.

#### 4.2. Non-generative methods

Unlike generative models, non-generative models do not explicitly represent the underlying data distribution; instead, they concentrate on learning direct mappings from input data to output predictions. Their approach is often more straightforward, typically requires fewer parameters, and generally demands less computing power. Below, we discuss several non-generative techniques: **Graph Attention Networks (GATs)**, **Graph Convolutional Networks (GCNs)**, **Autoencoders**, and **Feedforward Neural Networks (FNNs)**.

Graph Attention Networks (GATs), like GREMI [74], perform feature selection to identify significant variables and construct co-functional networks for each omics layer, enabling graph-based learning tailored to disease characterization. Similarly, MODILM [75] uses GATs to learn sample-specific and intra-association features from similarity networks for each omics modality. By building these similarity networks using cosine similarity, MODILM enhances the capture of omics-specific patterns, allowing better integration of multi-omics data and improving classification accuracy for complex diseases. MOGAT [76] mitigates bias by including only samples present across all eight modalities (mRNA,

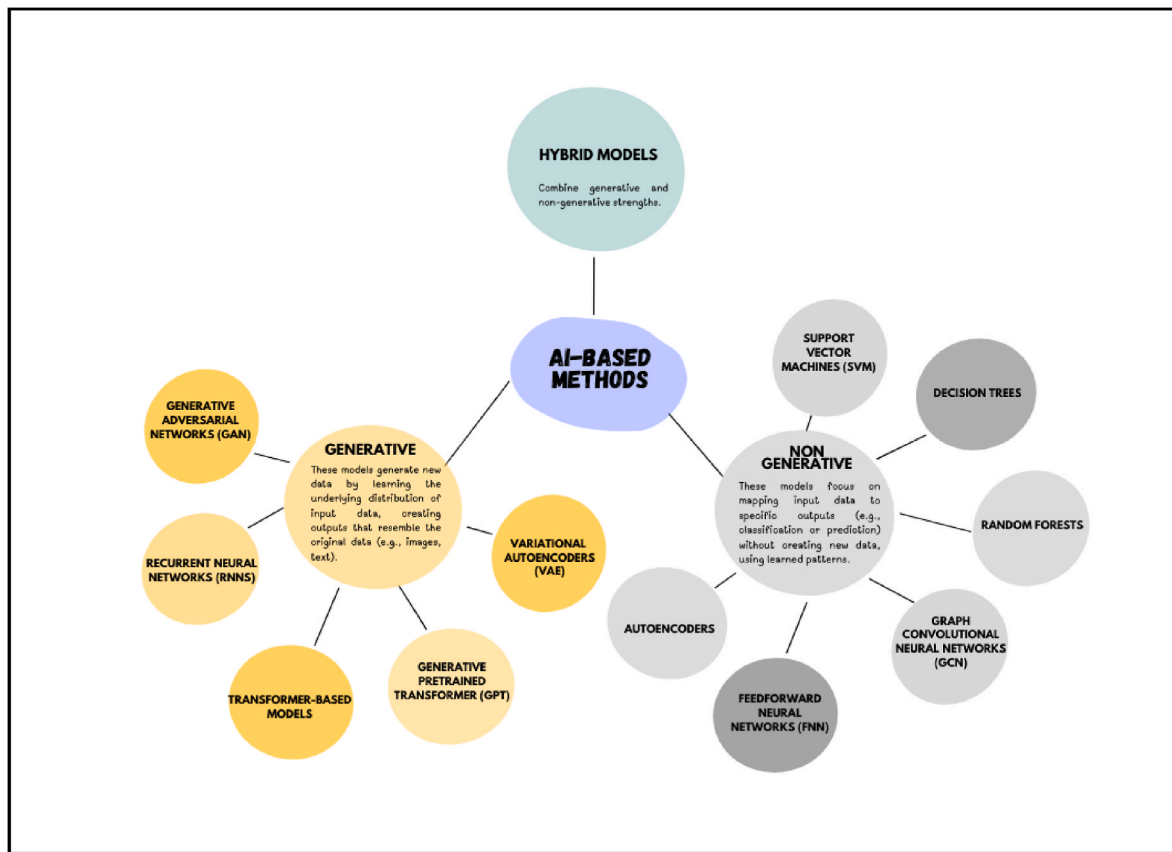


Fig. 5. Types of AI-based methods.

miRNA, lncRNA, methylation, SNV, CNA, eigengenes, clinical), with data cleaned and normalized to balance contributions.

Graph Convolutional Networks (GCNs), such as MOGONET [78], filter DNA methylation probes linked to gene promoters and apply variance thresholds (0.1 for mRNA, 0.001 for methylation) to retain informative features. cAGCN [80] normalizes gene expression, methylation, and CNV data using min-max scaling across global, gene-wise, and sample-wise dimensions, then maps these values onto a STRING PPI network for attention-based integration.

Autoencoders like MOCSS [85] process mRNA, miRNA, and methylation data using min-max normalization to [0,1], while Song et al.'s model [86] removes probes or genes with >50 % missing values and imputes gaps using the R impute package. Feedforward Neural Networks (FNNs), such as DeepKEGG [55], preprocess SNV data by binarizing mutations and normalize mRNA/miRNA expression mapped to KEGG pathways. DeepOmix [91] retains the top 5000 variable methylation and CNV features based on standard deviation and normalizes RNA-seq counts using DESeq2.

Convolutional Neural Networks (CNNs), like PCA-SMOTE-CNN [92], preprocess RNA-Seq, miRNA, and methylation data by filtering low-expression genes, applying DESeq2 or LIMMA for differential analysis, and using PCA for dimensionality reduction. Pathomic Fusion [93] extracts feature from histology images using CNNs, models cell relationships with GCNs and normalizes genomic data to z-scores before fusing them in a multimodal architecture.

In addition to purely generative and non-generative models, hybrid approaches have emerged as powerful solutions for multi-omics data integration. These models combine the strengths of both paradigms, leveraging the data generation and imputation capabilities of generative models while retaining the interpretability and efficiency of non-generative methods. For example, MultiGATAE [77] integrates GATs with autoencoders to achieve robust feature extraction and

classification. Similarly, SMMSN [81] combines GCNs with stacked autoencoders (SAEs) to enhance multi-omics data integration and improve cancer subtype classification. Hybrid models are particularly effective in addressing challenges such as missing data, high dimensionality, and modality heterogeneity, making them a promising avenue for future research in multi-omics oncology (see Table 4).

In addition, a recent survey by Lan et al. [95] critically examines transformer-based single-cell language models such as scGPT, emphasizing their role in enabling cross-modal data integration, scalability, and attention-driven interpretability in single-cell multi-omics analysis. While the focus is on single-cell resolution, the architectural innovations discussed also prove valuable in bulk-tissue oncology, particularly latent space regularization and attention-based weighting. These principles are reflected in recent models like TMO-Net, which adapts cross-modal fusion to infer interactions between mutation and expression profiles, and PATH-GPTOMIC, which repurposes the scGPT backbone for bulk RNA-seq by incorporating smoothing modules to mitigate distribution shifts across modalities.

In summary, common preprocessing steps include missing data handling (kNN or mean imputation), normalization (z-score, min-max), and feature selection (variance thresholds, ANOVA). Generative models like VAEs and GANs often automate preprocessing within their architectures, while non-generative tools rely on explicit pipelines. Despite advancements, challenges like batch effect correction and transparency in automated steps remain, highlighting the need for standardized workflows in future research.

#### 4.3. Technical foundations of multi-omics integration models

We outline three representative architectures for multi-omics integration (MOGONET, TMO-Net, and MultiGATAE) showing their distinct fusion mechanisms and loss designs for classification, prediction, and

**Table 3**  
DL-based methods for multi-omics data.

Model	Method	Modalities	Use-case	Year	Performance Metrics
OmicsGAN [63]	GAN	mRNA, microRNA expression	Cancer subtype classification & Survival prediction	2022	Demonstrates improved cancer prediction across BRCA, LUAD, and OV. Synthetic data significantly outperforms original and concatenated mRNA + miRNA in Area Under the Curve (AUC) (ER status AUC: 0.948 vs. 0.913; $P < 0.001$ ). Also identifies a greater number of significant predictive features, validating the utility of biologically informed integration.
Subtype-GAN [64]	GAN	mRNA, miRNA, CNV, DNA methylation	Cancer subtype classification	2021	Demonstrates significant survival separation ( $P = 5e-3$ ) and clinical parameter enrichment; recovers TCGA BRCA subtypes ( $P = 1e-7$ ); integrates multi-omics data across 10 cancers with superior performance to AE, iCluster, and LRAcluster.
ctGAN [65]	GAN	RNA-seq (mRNA expression)	Survival prediction	2024	Demonstrates significant Concordance Index (C-index) and log-rank p-value improvements over real data and other models in 9/11 cancers; up to ~15.7 % C-index increase (COAD); Outperformed trVAE and stVAE; validated by extensive cross-validation.
Al-Hurani et al. [66]	GAN + AE	mRNA, DNA methylation, miRNA, clinical	Subtype classification	2024	BRCA: Accuracy 95.1 %, AUC 1.00, Precision 100 %, Recall 81.5 %, F1 89.8 %; all outperform SMOTE-SVM-RBF. BLCA: Accuracy 88.8 %, AUC 91.2 %, Precision 85.4 %, Recall 75.9 %, F1 80.4 %; outperform NMF-GA except recall.
MetaCancer [67]	VAE	RNA-Seq, microRNA-Seq, DNA methylation	Subtype classification	2021	CVAE-based feature extraction outperforms network- and rank-based methods (83.8 % accuracy vs. 78.5 % and 74.2 %). MetaCancer multi-omics integration (88.9 % accuracy, 91.2 AUC) surpasses single-omics and ensemble SVM (82.5 % accuracy). mRNA contributes most among single omics.
MoVAE [68]	VAE	DNA methylation, miRNA, mRNA, CNV	Subtype classification	2024	Pan-Cancer single-omics best with mRNA (Acc 89.45 %, AUC 98.14 %). Two-omics avg 85.87 %, best with miRNA + mRNA (86.77 %). Three-omics avg 85.67 %, best at 87.55 %. Four-omics lower (79.76 %). Outperforms OmiVAE, X-VAE, ConsVAE, ConVAE; reconstructs missing omics with low error.
OmiEmbed [69]	VAE	Gene expression, DNA methylation, miRNA	Subtype classification	2021	Learns unified multi-omics embeddings for dimensionality reduction, classification, regression, and survival. BTM subtype: F1 0.832, Acc 0.875, AUC 0.994. GDC subtype: F1 0.968, Acc 0.977, AUC 0.999 with gene + DNA methylation + miRNA). Phenotypes: Disease stage F1 0.817, Primary site F1 0.972, Gender F1 0.956. Age prediction: MAE 8.37, RMSE 10.66, $R^2$ 0.479. Survival: C-index 0.772, IBS 0.166. Multi-task learning improves all metrics (C-index 0.782, F1 0.965, RMSE 10.63, MAE 6.68).
TMO-net [70]	VAE	Mutation, mRNA, CNV, DNA methylation	Subtype classification	2024	Learns pan-cancer multi-omics embeddings for subtype, prognosis, drug response, and metastasis. Pan-cancer subtype F1 0.751; METABRIC breast subtypes F1 0.921; metastasis F1 0.8980; drug response best AUC 0.697; prognosis avg C-index 0.6344. Excels at reconstructing missing omics and identifying key biomarkers.
cXVAE [71]	VAE	mRNA, DNA methylation	Disease subtype clustering	2024	Achieves best deconfounding in linear (ARI 0.712), nonlinear (ARI 0.646), and categorical confounders (ARI 0.664). Handles multiple confounders with ARI 0.634. Confounder ARIs $\leq 0.001$ . Recovers biological clusters masked by confounders. Outperforms other XVAE variants in accuracy and robustness.
PATH-GPTOMIC [72]	GPT	CNV, RNA-seq, Pathology images	Survival prediction	2024	Achieves top survival prediction on TCGA-GBMLGG (C-index 0.848) and TCGA-KIRC (C-index 0.754), surpassing PathOmics and Pathomic Fusion. Mix-up smoothing and gradient modulation improve scGPT embeddings for bulk RNA-seq integration.
mosGraphGPT [73]	GraphGPT	DNA methylation, Reverse Phase Protein Array (RPPA), mutations, RNA-seq, clinical	Classification & target discovery	2024	Achieves ~75.1 % accuracy in AD prediction vs. GNN, GAT, GIN, and UniMP. Learns signaling subnetworks revealing AD-specific pathways. Identifies top biomarkers via p-values and pathway enrichment. Excels in reconstructing signaling flows and extracting interpretable disease mechanisms.
GREMI [74]	GAT	mRNA, DNA methylation, miRNA	Subtype classification	2024	Outperforms 14 methods across ROSMAP, BRCA, LGG, KIPAN datasets. Achieves AD accuracy up to 86.4 %, BRCA subtype F1-weighted up to 87.7 %. Excels in multi-omics fusion, identifying disease modules linked to phosphoinositide signaling (AD) and Wnt signaling (BRCA). Validates LGG biomarkers in CGGA cohort (F1 = 0.731, $P = 4.8e-2$ ).
MODILM [75]	GAT	miRNA, mRNA, DNA methylation	Complex disease classification	2023	Outperforms MOMA, MOGONET, and other methods on ROSMAP, LGG, BRCA, SKCM, LUSC datasets, with up to +21.6 % gains in F1-weighted. Best performance with 2 GAT layers and 2 MLP layers. Ablations show GAT and VCDN crucial for accuracy. Multi-omics integration improves results, though mRNA alone can sometimes match triple-omics.
MOGAT [76]	GAT	mRNA, DNA methylation, miRNA, CNV, mutations, RPPA, lncRNA, clinical,	Subtype classification & survival analysis	2024	Outperforms MOGONET and SUPREME on BRCA subtype prediction. Achieves macro-F1 0.826 (all 8 omics), with mRNA contributing most. GAT embeddings improve survival

(continued on next page)

Table 3 (continued)

Model	Method	Modalities	Use-case	Year	Performance Metrics
MultiGATAE [77]	GAT + AE	mRNA, DNA methylation, miRNA	Subtype classification	2022	stratification (log-rank P 2.10e-30, HR 0.10) vs. raw features. Visualizations show clearer subtype clusters than raw data. Outperforms eight clustering methods on eight TCGA cancers (KIRC, BRCA, COAD, SKCM, GBM, LUSC, LIHC, OV), achieving highest negative log10 p-values and C-index on most datasets. Identifies prognostic subtypes with distinct survival curves. Multi-omics integration improves results over single-omics, with DNA methylation most informative.
MOGONET [78]	GCN	mRNA, DNA methylation, miRNA	Subtype classification	2021	Outperforms nine supervised multi-omics methods on ROSMAP (AD), LGG, BRCA, and KIPAN datasets. GCNs learn omics-specific relations; VCDN captures cross-omics label correlations. Multi-omics integration boosts performance over single omics. Identifies biomarkers linked to AD and breast cancer pathways. Ablations show GCN and VCDN critical for accuracy.
MoGCN [79]	GCN	CNV, RNA-seq, RPPA	Subtype classification	2022	Outperforms DT, KNN, RF, SVM, DNN, GrassmannCluster, and HOPES on BRCA and KIPAN datasets. Achieves BRCA subtype identification and 97.7 % accuracy/F1 in KIPAN classification. Integrates AE features and SNF-based patient similarity networks via GCN for stable, interpretable results. Discovers pathways including Wnt, PI3K-AKT-mTOR, EMT processes.
cAGCN [80]	GCN	mRNA, DNA methylation, CNV	Breast cancer subtyping	2023	Outperforms simple GCN, MLPs, and other methods on BRCA, COAD, and Pan-cancer datasets. Achieves AUC up to 0.97 and robust performance despite class imbalance, particularly excelling in basal-like BRCA classification. Uses SE and cross-omics attention for feature fusion and interpretability. Identifies subtype-specific driver genes and pathways and maintains predictive performance on unlabeled samples.
SMMSN [81]	GCN + SAE	mRNA, DNA methylation, miRNA	Cancer clustering	2024	Outperforms traditional and deep clustering methods on labeled and unlabeled datasets. Achieves 85.34 % ACC on KIPAN and lowest survival p-values across five cancers. Dual self-supervised learning integrates SAE and GCN for robust multi-omics fusion. Identifies subtype-specific drivers (EGFR in GBM, Wnt pathway in BIC).
DeepMOCCA [82]	GCN	differential gene expression, differential methylation, CNVs, SNVs, clinical	Survival prediction	2021	Integrates multi-omics data in a graph-CNN framework to predict patient survival directly from tumor-specific molecular and clinical features. Incorporates a sample-specific graph-attention module to highlight prognostic genes, many of which correspond to established cancer drivers. Demonstrates that learned representations naturally group related tumor types (COAD with READ, LGG with GBM), underscoring the model's ability to capture biologically meaningful patterns.
Braytee et al. [83]	AE	CNV, DNA methylation, miRNA, RNA-seq, clinical	Cancer risk clustering & survival analysis	2024	Combines CANDECOMP/PARAFAC (CP) tensor decomposition with autoencoders to extract latent factors from multi-omics data, followed by SHAP for biomarker interpretation. Outperforms MOFA in stratifying glioma and breast cancer patients into risk groups based on survival, achieving significant p-values in Kaplan–Meier analysis. Identifies subtype-relevant biomarkers and pathways (immune signaling in glioma, cytoskeleton organization in breast). Latent features enable tumor purity classification in breast cancer with up to 0.69 accuracy via Random Forest.
DeepAutoGlioma [84]	AE	RNA-seq (gene expression), DNA methylation	Glioma subtyping	2023	Integrates survival-associated DEGs and promoter DMRs via an autoencoder to generate biologically grounded latent features. A 1D CNN achieves 98.03 % accuracy for LGG and 94.07 % for GBM subtyping, significantly outperforming both single-omics and random-feature baselines. External validation yields 95.23 % (LGG) and 90.26 % (GBM) accuracy. Identified latent features map to pathways such as ALK signaling and cell–cell adhesion, underscoring the value of biologically guided feature selection and di-omics integration for high-accuracy classification.
MOCSS [85]	AE	mRNA, miRNA, DNA methylation	Cancer subtyping	2023	Learns multi-omics representations via autoencoders and contrastive learning with orthogonality constraints. Outperforms SNF, NEMO, DeFusion, MDICC, Subtype-GAN, and Subtype-DCC across five cancers, achieving top clustering accuracy and survival stratification. Identifies subtype-specific biomarkers (SLC14A2, miR-4746-5p in LUAD) and reveals clinical correlations such as smoking-linked aggressive subtypes.
Song et al. [86]	AE	mRNA, miRNA, DNA methylation	Colorectal cancer survival stratification	2022	Integrates multi-omics data via autoencoders to identify two CRC risk groups (G1, G2), with G2 linked to poorer survival. Outperforms PCA, t-SNE, NMF, and individual Cox models in capturing survival-related features. Validated across five external cohorts. Reveals distinct DEGs, DEmiRNAs, and DMGs between subtypes, implicating pathways such as Wnt, PI3K-Akt, and immune regulation in aggressive CRC biology.

(continued on next page)

Table 3 (continued)

Model	Method	Modalities	Use-case	Year	Performance Metrics
DeepMoIC [87]	AE + GCN	mRNA, DNA methylation, miRNA, CNV, RPPA	Cancer subtyping & survival prediction	2024	Combines autoencoders with deep GCNs and a patient similarity network to classify subtypes and predict survival across BRCA, KIPAN, LGG, and TCGA-pan-cancer cohorts. Reports BRCA accuracy $\approx 84.3\%$ and F1 $\approx 81.0\%$ , surpassing DeepMO, MOGONET, and MoGCN. Autoencoder-selected biomarkers are enriched in pathways such as p53 and Wnt. Kaplan–Meier curves demonstrate significant subtype separation (log-rank $p < 0.01$ ). Ablation studies on GCN depth, residual connections, and the PSN confirm robustness.
CustOmics [88]	VAE	CNVs, Gene expression, DNA methylation	Cancer classification & survival analysis	2023	Introduces a variational autoencoder for multi-omics integration, outperforming MFA, UMAP, and NMF in pan-cancer tumor classification and survival prediction. Achieves superior accuracy and survival stratification, leveraging cross-omics interactions. SHAP-based interpretability reveals subtype-relevant genes like <i>TFPI</i> . Late and joint integration improve multi-omics learning. RNA-Seq is the most informative modality. For survival analysis, CustOmics significantly outperforms other methods across five TCGA cohorts, successfully stratifying patients into high/low-risk groups
DeepKEGG [55]	FFN + AE + attention	SNV, mRNA, miRNA, clinical	Cancer recurrence prediction	2024	Leverages a pathway self-attention mechanism integrated with autoencoders to capture gene/miRNA-pathway relations in multi-omics data. Outperforms state-of-the-art methods (MOGONET, PathCNN ...) in AUC and AUPR across multiple TCGA and TARGET datasets, showing up to 6–7% gains. Identifies potential biomarkers and pathways (MAPK, Hippo, PI3K-Akt ...) linked to cancer recurrence. Demonstrates superior performance over single-omics models and provides biologically interpretable insights for personalized cancer prediction
DeepMO [89]	DNN (late fusion)	mRNA, DNA methylation, CNV	Breast cancer subtyping	2020	Outperforms single-omics models and other integration methods (MKL, Elastic Net, RF), achieving mean AUC up to $\sim 0.96$ in binary breast cancer subtype classification and stable multi-class performance. Incorporates chi-squared feature selection to improve accuracy and training speed. Identifies subtype-relevant genes and pathways (MED1, GRB7, cell cycle, morphogenesis ...).
DSCCN [90]	DNN + Attention	mRNA, DNA methylation	Breast cancer subtyping	2024	Integrates differential analysis and Sparse Canonical Correlation Analysis (FGL-SCCA) to identify correlated mRNA and DNA methylation features, followed by a deep neural network with module encoding and attention for subtype classification. Achieves superior performance in both binary and multi-class breast cancer subtype prediction, outperforming methods like DIABLO, SMSPL, DeepMO, and ensemble classifiers. Identifies biologically relevant gene signatures linked to subtypes and offers insights into regulatory correlations between omics layers.
DeepOmix [91]	FFN	Mutations, CNA, gene expression, DNA methylation	Survival prediction	2021	Integrates multi-omics data and prior biological knowledge via a feed-forward neural network with a functional module (pathway) layer representing KEGG and Reactome pathways. Learns non-linear, low-dimensional pathway embeddings to predict survival outcomes. Outperforms methods like glmboost, IPF-LASSO, block forest, DeepHIT, and DeepSurv in C-index across eight TCGA cancer types. Enables functional interpretation by linking pathway activations to survival risk, highlighting pathways associated with high- and low-risk patient groups.
PCA-SMOTE-CNN [92]	CNN	RNA-Seq, miRNA, DNA methylation	Lung cancer staging	2024	Employs PCA for dimensionality reduction and SMOTE for class balancing, feeding integrated multi-omics data into a 1D CNN with convolutional, pooling, and dense layers for multi-class lung cancer stage prediction. Achieves high performance (accuracy and F1-score of 0.97) and outperforms methods like LungDWM, CVAE-based models, CC2DT, SVM ensembles, and classical ML algorithms. Demonstrates superior results over single-omics inputs and reaches a perfect AUC (1.00) when integrating all three omics types, highlighting the advantage of multi-omics integration.
Pathomic Fusion [93]	CNN-GCN-SNN	CNV, mutation, RNA-Seq, Histology, Cell graphs	Survival prediction	2022	Integrates histology images, cell graphs, and genomic features using multimodal deep learning with late tensor fusion. Combines CNNs for image ROIs, GCNs for cell graphs, and SNNs for genomic data. Achieves higher concordance index (glioma: 0.826; CCRCC: 0.720) than unimodal models and prior state-of-the-art, improving survival prediction and grade classification. Provides multimodal interpretability linking morphological regions, cellular structures, and genomic markers (IDH1, PTEN, EGFR in glioma; CYP3A7, PITX2 in CCRCC) to patient risk. Outperforms WHO grading and traditional Cox models in patient stratification.

**Table 4**  
Comparison of generative, non-generative, and hybrid models in multi-omics data integration.

Aspect	Generative	Non-Generative	Hybrid
<b>Characteristics</b>	Learn data distributions, generate new data.	Direct input-output mapping.	Combines both approaches
<b>Examples</b>	VAEs (MoVAE, TMO-Net), GANs (ctGAN, Subtype-GAN), GPTs (mosGraphGPT).	GATs (GREMI, MOGAT), GCNs (MOGONET, cAGCN), Autoencoders (MOCSS, DeepAutoGlioma), FNNs (DeepKEGG).	MultiGATAE (GAT + AE), SMMSN (GCN + SAE)
<b>Main Uses</b>	Latent feature learning, denoising, augmentation, imputation, cross-modal alignment, downstream classification and survival	Feature extraction, classification, clustering, subtype/risk prediction.	More robust multi-modal fusion, partial-modality training, subtype discovery.
<b>Strengths</b>	Handles missing data and high dimensionality.	Simpler, efficient, interpretable.	Combines strengths of both
<b>Limits</b>	Computationally intensive and less interpretable.	Limited ability to handle missing data.	More complex to implement and tune.
<b>Missing Data</b>	Can reconstruct and impute from latent.	Needs external imputation and masking.	Partial-modality training and shared latent mitigate missing views.
<b>Interpretability</b>	Latent factor analysis, reconstruction influence, and pathway enrichment.	Saliency, attention, SHAP; sometimes clearer due to direct mapping.	Combined latent + attention + attribution

subtype discovery.

MOGONET exemplifies a dual-phase graph-based approach for supervised multi-omics integration. In the first phase, each omics modality is represented as a graph using sample-wise cosine similarity, followed by graph convolutional networks (GCNs) to learn omics-specific embeddings and predictions. The adjacency matrix is normalized symmetrically as  $\hat{A} = D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}}$ , where  $\tilde{A}$  is the adjacency matrix with self-loops and  $D$  is the diagonal node degree matrix. Each GCN is optimized using a cross-entropy classification loss. In the second phase, MOGONET captures cross-omics interactions by computing an outer product of predicted class probabilities across modalities, forming a discovery tensor. This tensor is reshaped and input into a View Correlation Discovery Network (VCDN), which integrates multimodal information for final classification. The overall objective combines modality-specific classification losses with the VCDN loss.

TMO-Net models multi-omics data  $X = \{X_1, X_2, \dots, X_m\}$ , where each modality  $X_i \in \mathbb{R}^{n \times d_i}$  represents  $n$  samples and  $d_i$  features. Each modality is encoded using a modality-specific variational autoencoder (VAE) to generate latent embeddings  $z_i$ . These embeddings are optimized through evidence lower bound (ELBO) objective that includes both self-modal and cross-modal reconstruction losses, denoted as  $L_{ELBO}$ . A Cross Fusion Module (CFM) integrates  $\{z_1, \dots, z_m\}$  using a Product-of-Experts mechanism, supporting inference when some modalities are missing. The fused representation is used for downstream tasks including classification (cross-entropy loss  $L_{cls}$ ), survival prediction (Cox loss  $L_{cox}$ ), and drug response prediction (mean squared error loss  $L_{mse}$ ) when included. These are collectively denoted as  $L_{task}$ , selected based on the prediction objective. The total training loss is defined as  $L_{total} = L_{ELBO} + \lambda_1 \cdot L_{cls} +$

$\lambda_2 \cdot L_{cox} + \lambda_3 \cdot L_{mse}$ , where  $\lambda_1, \lambda_2$ , and  $\lambda_3$  weight the auxiliary losses (unused tasks have  $\lambda = 0$ ). This formulation allows TMO-Net to learn informative, aligned, and modality-agnostic representations, while supporting robust prediction across heterogeneous and partially missing multi-omics datasets.

MultiGATAE performs unsupervised cancer subtyping through integrated graph representation learning. First, it constructs fused similarity graphs across omics modalities using Similarity Network Fusion (SNF). Initial similarity matrices are computed using a kernel-based distance metric. Cross-omics fusion is achieved through iterative similarity network updates. A Graph Attention Autoencoder then processes the fused graph, where attention coefficients determine neighbor importance during graph aggregation. The model minimizes adjacency reconstruction loss (for example  $L_{recon} = \|A - \hat{A}\|_F^2$  where  $\hat{A}_{ij}$  is typically derived from latent embeddings such as  $\hat{A}_{ij} = \sigma(z_i^T z_j)$ ). After obtaining modality-specific latent representations, MultiGATAE fuses them using attention-based weights to capture cross-modal importance. The fused representation  $Z_{fuse}$  often computed as a weighted sum  $Z_{fuse} = \sum_{v=1}^m \beta^{(v)} Z^{(v)}$ , where  $Z^{(v)}$  denotes the embedding from modality  $v$ , and  $\beta^{(v)}$  are learnable attention weights reflecting the relative importance of each modality. K-means clustering is then applied to  $Z_{fuse}$  to identify biologically meaningful cancer subtypes. This unsupervised pipeline enables MultiGATAE to capture nonlinear cross-omics interactions and identify biologically meaningful cancer subtypes.

The technical architecture presented demonstrates diverse yet powerful deep learning strategies for multi-omics integration. Key innovations include: MOGONET's modality-specific GCNs followed by cross-omics discovery tensors and a View Correlation Discovery Network (VCDN) for supervised classification; TMO-Net's self-modal and cross-modal variational autoencoders aligned via a Product-of-Experts fusion module to enable robust inference when modalities are missing; and MultiGATAE's Similarity Network Fusion of sample graphs coupled with an attention-based graph autoencoder for unsupervised subtype clustering. These models highlight the centrality of representation learning, fusion mechanisms, and task-tailored architectures in overcoming multi-omics heterogeneity.

Each approach addresses core integration challenges: MOGONET smooths sparse signals via graph propagation and leverages cross-omic label correlations in VCDN; TMO-Net reconstructs missing modalities through cross-modal VAEs; and MultiGATAE reduces dimensionality and enhances interpretability via graph attention. All three consistently outperform single-omics baselines on their respective classification or clustering tasks. Nonetheless, issues such as batch-effect correction, scaling to larger cohorts, and standardized benchmarking persist, underscoring the need for ongoing methodological refinement (see Table 5).

## 5. Critical research gaps and validation challenges

This review identified several critical limitations in the current landscape of deep learning models for multi-omics integration, directly impacting their translational pathway. Protecting generalizability and ensuring real-world utility are tightly related to overcoming data constraints and validation gaps.

### 5.1. Predominant reliance on TCGA-derived data

Even though this review intentionally focused on studies using TCGA data, a key finding is the field's overwhelming dependence on this single-source dataset. While TCGA offers high-quality molecular profiles, this reliance introduces important concerns regarding ecological validity. Notable limitations include the absence of multi-institutional Electronic Health Record (EHR) integration, lack of longitudinal molecular follow-up, and limited demographic diversity in the TCGA-

**Table 5**

Comparative analysis of methodological robustness in MOGONET, TMO-Net, and MultiGATAE.

Challenge	MOGONET	TMO-Net	MultiGATAE
<b>Sparsity Handling</b>	Graph propagation across modality-specific sparse graphs	Product-of-Experts (PoE) fusion + self/cross-modal reconstruction support missing modalities	GAT-based attention focuses on informative edges in fused graphs
<b>Bias Mitigation</b>	Modality-specific GCNs + VCDN	Shared VAE latent space + PoE align modalities	SNF + attention AE balance modality contributions
<b>Dimensionality Reduction</b>	GCN embedding layer	VAE encoder latent space	Autoencoder latent space
<b>Interpretability</b>	VCDN class probability tensor (limited feature saliency)	Gradient/perturbation saliency on latent or task heads	Attention coefficients
<b>Computational Efficiency</b>	Separate GCN per modality + VCDN (moderate)	Parallel per-modality VAEs + lightweight PoE fusion	Lightweight GAT + decoder
<b>Validation Rigor</b>	Cross-Validation on METABRIC, subtype prediction	Subtype classification, survival, drug response, missing-modality ablations	Unsupervised subtype discovery & survival analysis
<b>Clinical Translation</b>	ER/HER2/PAM50 subtype classification (experimental)	Pan-cancer subtype & prognostic modeling (experimental)	Subtype discovery & biomarker enrichment (experimental)

PanCancer cohort [96–98]. This imbalance risks reduced model performance in minority populations; in many cases, model performance for underrepresented groups remains unknown.

As a result, reported high performance metrics may not generalize well to broader, more heterogeneous clinical populations.

This imbalance risks reduced model performance in minority populations; in many cases, model performance for underrepresented groups remains unknown.

Emerging research highlights how site-specific technical artifacts in TCGA data can bias downstream AI applications. For example, Dehkharghanian et al. (2023) [99] showed that deep learning models trained on TCGA histopathology slides could predict the image acquisition site with up to 86 % accuracy, using non-biological cues such as staining protocols, scanner configurations, and other institution-specific artifacts. These medically irrelevant features risk confounding model interpretation, particularly in histopathology-based cancer classification and retrieval tasks. Such findings underscore the need for caution and artifact mitigation when using centralized datasets like TCGA.

This overreliance also narrows multi-omics integration efforts essential for decoding the layered complexity of cancer biology. Although TCGA offers valuable infrastructure, it represents only a subset of available resources. As Das et al. (2020) [100] emphasize, a more generalizable approach requires leveraging a broader ecosystem of omics repositories, including the International Cancer Genome Consortium (ICGC), Catalogue Of Somatic Mutations In Cancer (COSMIC), The Pathology Atlas, Gene Expression Omnibus (GEO), and PRoteomics IDentifications (PRIDE). While integrating such diverse datasets introduces novel insights, it also presents challenges related to heterogeneity, standardization, and interpretability.

A key challenge ahead is developing methods to harmonize multi-omics data without overfitting to biases from a single source like TCGA which is an essential step toward realizing the full potential of precision oncology.

## 5.2. Scarcity of clinical validation and workflow integration

Despite excelling in clustering and subtype prediction, these models have yet to be integrated into clinical workflows or therapeutic protocols. Although limited external validation was performed by some models such as using ICGC and CGGA, none demonstrated prospective clinical application, interoperability with clinical data systems, or real-world validation [101]. This gap underscores a critical issue: the rapid pace of architectural innovation is not matched by translational rigor. As model complexity grows, so too does the risk of clinical irrelevance. Without attention to usability, interpretability, and regulatory alignment, even the most advanced models are unlikely to earn clinician trust or reach patients. The promise of precision oncology is thus undermined not by performance limitations, but by a persistent failure to align outputs with actionable clinical endpoints.

Even among the most cited systems, real-world deployment remains absent. This observation aligns with a broader trend in medical AI literature, where the majority of developed tools fail to reach clinical adoption, primarily due to reproducibility challenges, limited generalizability, and regulatory barriers [102,103].

Ultimately, without a clear focus on real-world use, clinical needs, and regulations, even the best models will stay out of everyday cancer care. To bridge the gap between methodological innovation and clinical utility, future research in multi-omics data integration must prioritize generalizability, fairness, and implementation. While deep learning models have demonstrated impressive predictive power, their real-world impact remains limited by dataset biases, lack of interpretability, and minimal clinical validation. Addressing these challenges will require standardized benchmarks beyond TCGA, integration with diverse patient cohorts and electronic health records, and greater emphasis on transparent, reproducible modeling practices. Importantly, the field must shift from purely performance-driven evaluations toward actionable outcomes that support precision oncology in everyday clinical decision-making.

## 6. Addressing challenges and opportunities in multi-omics oncology

Every approach mentioned earlier aims to tackle the challenges of multi-omics data by employing suitable preprocessing and preparation techniques tailored to specific problems. There are several common strategies that are regularly employed in these methods. Additionally, some approaches are especially well-aligned with the underlying framework, improving the overall efficiency and effectiveness of the integration process.

Looking more closely into how these challenges are addressed, we can outline a general workflow for handling data modalities such as those derived from the TCGA database. This includes strategies to ensure effective transparency, explainability, and ethical handling of medical data. Furthermore, potential future pathways in the domain of multi-omics integration are emerging, leading to more thorough and effective multi-modality analyses aimed at improving precision in medical decisions and enhancing healthcare systems.

### 6.1. Preprocessing multi omics data by modality

Proper preprocessing is essential to guarantee data quality and compatibility. This involves addressing incomplete data, noise, and high dimensionality, with techniques tailored to each modality's unique traits.

**RNA-seq:** Preprocessing begins with filtering low-variance genes (standard deviation is often  $<1$ ) or genes with very low overall expression, as these contribute little to downstream analyses. Missing values are commonly imputed using techniques such as k-nearest neighbors or nearest-neighbor averaging. Many deep learning models, including ctGAN [65] and MetaCancer [67], rely on learned RNA-seq

embeddings to represent complex gene expression patterns. In the context of bulk RNA-seq data, additional steps like embedding-based denoising or smoothing, as adapted from transformer-based models such as scGPT [94], are often applied to improve generalizability and reduce noise.

**DNA Methylation:** Due to the inherently sparse and noisy nature of methylation datasets, it is common practice to exclude probes exhibiting a high proportion of missing values (typically ranging from 20 % up to 90 %) or those with low beta values. Dimensionality reduction techniques are often employed to retain probes displaying substantial variability, or to select survival-associated markers using supervised approaches such as Cox proportional hazards (CoxPH) models. Imputation strategies may involve simple averaging or advanced latent representation methods like variational autoencoders (VAEs), as implemented in studies such as MoVAE [68] and TMO-Net [70]. Tools like ELMER also facilitate the reconstruction of gene regulatory networks from methylation data.

**CNVs:** Copy number segments that contain substantial missingness (>20 %), display low mean absolute values (<0.20), or have limited coefficients of variation (<0.20) are typically filtered out. Integration of CNVs with other genomic features can leverage clustering algorithms shared nearest neighbor (sNN) graph clustering strategy, as in PATH-GPTOMIC [72].

**miRNA-Seq:** Lowly expressed or poorly annotated miRNA sequences are generally excluded from analysis, and missing entries are handled using statistical imputation or latent-space modeling. It is standard for miRNA-seq data to be co-analyzed with RNA-seq as part of multi-omics integration pipelines.

**Cross-Modality Challenges:** The complexity of missing data in multi-omics poses significant hurdles. For instance, true zeros in gene expression are often misinterpreted, requiring further investigation. Tools like ComBat or Harmony address batch effects, while VAEs assist in dimensionality reduction and modality-specific representations.

Summing up, common preprocessing steps such as missing data handling (kNN, VAEs), normalization, and feature selection are tailored to each modality's needs. Generative models automate many steps within their architectures, while non-generative tools rely on explicit pipelines. Despite these advancements, challenges like batch effect correction and transparency in automated workflows persist. For example, distinguishing true zeros from missing values in gene expression remains difficult and requires further research.

## 6.2. Pathological images fusion with biological data

Biomedical imaging refers to a variety of imaging techniques used in medicine, such as conventional X-rays and ultrasound. Obtaining medical images frequently requires expensive equipment and can be time consuming. The availability of datasets for thoracic computed tomography (CT), chest radiographs, and mammograms has accelerated research into lung diseases and breast cancer. In a study by Yao et al. [104] for example, CT-scans and transcriptomic data are integrated to investigate the relationship between CD38 expression and survival outcomes in epithelial ovarian cancer patients.

Whole-slide images (WSIs) contain detailed information on tissue structures and cell types, which can be used to determine the malignancy level and tumor development stage. As a result, pathologists frequently regard these pathological characteristics as critical criteria for cancer diagnosis and staging. While WSIs are valuable for assessing tumor morphology, they lack the molecular insights into biological processes, mutations, and pathways driving cancer progression. Integrating histological and genomic data overcomes these limitations by linking structural patterns to biological pathways represented as nodes in a genomic subgraph.

For survival analysis, traditional approaches frequently rely on patch-based multiple instance learning (MIL), treating each image patch as an independent instance, thereby neglecting critical spatial

relationships and morphological context. In contrast, the Pathology-Genome Heterogeneous Graph (PGHG) [105] introduced a graph-based approach that models histology image patches as nodes and defines edges based on their spatial adjacency. This design preserves the structural organization of tissues and enables the model to capture complex spatial interactions within the tumor microenvironment. By integrating these spatial dependencies with transcriptomic profiles through attention-based graph learning, PGHG enhances prognostic prediction and provides biologically interpretable insights into cancer progression.

The histology data in Pathomic Fusion [93] is processed using two methods. Firstly, CNNs are employed for extracting image-based features, while GCNs are used for capturing cell-based features. This enables the system to capture both the structure of the tissue and the interactions between cells. The extracted features from histological data, along with genomic data, are then integrated into a multimodal tensor that represents all possible interactions between different modalities. To enhance the accuracy of predictions, a gating-based attention mechanism is employed to highlight relevant features and reduce the importance of less significant ones.

## 6.3. Transparency and explainability of AI models

In precision medicine, AI-based decisions derived from multi-omics data integration must be transparent and interpretable. The lack of transparency in complex AI models like DL could lead to doubts in the medical fields. Ethical considerations must focus on developing interpretable models that allow stakeholders to understand how predictions are made and how they affect clinical decisions. This is particularly important when human health and personalized treatment plans are involved. Transparency and explainability are recognized as essential attributes of AI systems, influencing user expectations, cultural norms, and legal aspects [106,107].

Balasubramaniam et al. [5], proposed a framework to define explainability requirements for AI systems, addressing questions like *whom* and *what* to explain. In their recent study, Govea et al. [107] demonstrated that transparent recommended systems can combine advanced algorithms with explainability techniques such as using Local Interpretable Model-agnostic Explanations (LIME) and Shapley additive Explanations (SHAP) without sacrificing efficiency, enhancing trust and precision. For example, SHAP-improved deep learning models showed higher recommendation accuracy, proving the link between explainability and algorithmic effectiveness.

Researchers increasingly adopt explainable AI (XAI) to interpret omics data and uncover biological insights. XAI focuses on simplifying black-box models or providing post-hoc explanations (feature relevance) [108,109]. A systematic review of 405 studies [10] mapped XAI applications in omics, identifying patterns like integrating biological knowledge such as gene regulatory networks to improve plausibility. For instance, TMO-Net reveals molecular feature impacts on clinical outcomes using explainable deep learning. GREMI isolates disease-relevant subgraphs via Monte Carlo tree search. DeepMOCCA lacks interpretability for individual omics features such as methylation sites, highlighting the need for finer-resolution diagnostics.

Models like DeepOmix and DeepKEGG incorporate biological pathways and co-expression analyses to align outputs with real-world systems, improving interpretability. Graph convolutional networks (GCNs) leverage patient similarity networks (PSNs) to capture inter-patient relationships, enhancing subtype classification and survival prediction. For example, MOGONET and MoGCN prioritize preprocessing to ensure high-quality PSN inputs, boosting prediction accuracy. The Monte Carlo tree search (MCTS) [110] is also widely used for pinpointing significant features or subgraphs within datasets, as seen in GREMI, where it explores local-view subgraphs to identify modules contributing to disease characterization.

6.4. Data privacy and security in precision medicine

Multi-omics data often includes sensitive patient information, which raises privacy concerns. When integrated using AI for precision medicine, ensuring that patient data is anonymized and securely stored is crucial [111]. The ethical implications include risks of data breaches or misuse, especially when AI models are trained on a large scale, and data derived from multiple sources. In their paper, J. Zhou et al. [9] described various scenarios and strategies to protect patient privacy across distinct stages of omics data use and the development of AI-driven omics methods, starting with data control and leading up to model deployment and sharing. Privacy-preserving data mining techniques are designed to process anonymized data, enabling knowledge extraction and AI system development without compromising individual privacy.

Over the last five years, privacy protection regulations for health data have undergone significant changes across multiple regions. In the United States, HIPAA [112] continues to regulate protected health information (PHI), while additional laws such as California Consumer Privacy Act (CCPA) [113] improve individual control over health data by requiring consent, access, and deletion rights. The HITECH Act [114], enacted in 2009, promotes the adoption of electronic health records (EHRs) and strengthens HIPAA by introducing stricter enforcement of privacy and security rules related to PHI, particularly in digital formats.

In the European Union, The General Data Protection Regulation (GDPR) [115] sets guidelines for consent and gives individuals control over their personal information, including health data, and has had a

worldwide impact in shaping data protection. Similarly, Brazil's LGPD [116] and China's PIPL [117] mirror GDPR by imposing strict consent, breach notification, and cross-border data transfer restrictions on sensitive health data. Japan's APPI [118] and Australia's NDB Scheme [119] have strengthened their regulatory frameworks by mandating reporting requirements and prioritizing health information security. The UK Data Protection Act 2018 [120] incorporates GDPR principles, ensuring strong protection for health data after Brexit. New Zealand and Mexico have also improved health-specific data privacy regulations.

Worldwide, there is a strong focus on safeguarding health data through transparent regulations, ethical approvals, and robust security measures, particularly in research involving human genomic data.

Table 6 includes laws and regulations that are still in effect or have been revised within the last five years to protect sensitive health data. These regulations are critical when discussing AI and multi-omics data privacy to ensure the ethical handling and protection of sensitive personal and healthcare information.

6.5. AI bias and fairness in healthcare

One critical issue is the reliance on large-scale datasets such as TCGA, which suffer from demographic limitations despite their importance. For example, genotype-derived ancestry data for 9899 TCGA cases, as reported by Oak et al. (2018) [123], indicate that over 82 % are of European descent, while only around 9.8 % are African, 6.6 % are East Asian, and just 0.5 % are Native American/Latin American, leaving

**Table 6**  
Health data privacy protection regulations.

Regulation	Primarily Aspects	Target users	Enforcement Agency	Target Region
HIPAA [112]	<ul style="list-style-type: none"> <li>- Secure patient health information (PHI).</li> <li>- Regulate the use and disclosure of protected health information.</li> <li>- Give patients the right to access and correct their health information.</li> </ul>	Healthcare providers, insurers, and clearinghouses	U.S. Department of Health and Human Services (HHS)	USA
CCPA [113]	<ul style="list-style-type: none"> <li>- California residents have rights over their personal data, including sensitive health data. - This includes access, deletion, and opt-out from data sales.</li> <li>- Emphasize transparency and consumer control.</li> <li>- Apply to personal data, including health data unless exempt under HIPAA.</li> </ul>	Businesses processing data of California residents	California Attorney General	California, USA
HITECH Act [114]	<ul style="list-style-type: none"> <li>- Strengthen HIPAA.</li> <li>- Establish breach notification rules.</li> <li>- Concentrate on electronic health record (EHR) security.</li> </ul>	Healthcare providers, insurers, business associates	U.S. Department of Health and Human Services (HHS)	USA
GDPR [115]	<ul style="list-style-type: none"> <li>- Outline detailed data protection policies, including those governing sensitive health information.</li> <li>- Give individuals control over their personal data (such as the right to be forgotten, access, and portability).</li> <li>- Strict consent requirements for processing health data.</li> </ul>	Any organization processing personal data of EU citizens	National Data Protection Authorities in EU member states	European Union
Brazil LGPD [116]	<ul style="list-style-type: none"> <li>- Safeguard sensitive personal information, including health data.</li> <li>- Consent is required for processing.</li> <li>- The right to access, correct, and delete data.</li> </ul>	Organizations handling personal data	Brazilian National Data Protection Authority (ANPD)	Brazil
PIPL [117]	<ul style="list-style-type: none"> <li>- A comprehensive data protection law.</li> <li>- Explicit consent is required for processing health data.</li> <li>- Strict rules for cross-border data transfer.</li> </ul>	All entities processing personal data	Cyberspace Administration of China (CAC)	China
APPI [118]	<ul style="list-style-type: none"> <li>- Protect personal information, including health information.</li> <li>- Establish transparency and consent requirements.</li> <li>- Include cross-border data transfers.</li> </ul>	Organizations handling personal data in Japan	Personal Information Protection Commission (PPC)	Japan
NDB Scheme [119]	<ul style="list-style-type: none"> <li>- Mandatory reporting of data breaches, particularly for health-related breaches.</li> <li>- Provide guidance on how to handle sensitive health data.</li> </ul>	Organizations handling personal data	Office of the Australian Information Commissioner (OAIC)	Australia
Data Protection Act 2018 [120]	<ul style="list-style-type: none"> <li>- Implement GDPR and strengthen UK health data protection.</li> <li>- Establish standards for processing personal data and strengthen safeguards for sensitive data categories, such as health.</li> </ul>	Organizations handling personal data in the UK	Information Commissioner's Office (ICO)	United Kingdom
My Health Records Act 2012 [121]	<ul style="list-style-type: none"> <li>- Oversee the operation of the nationwide My Health Record system.</li> <li>- Maintain privacy and security controls for personal health information stored electronically.</li> <li>- Patients can access and manage their own medical records.</li> </ul>	Healthcare providers using the My Health Record system	Australian Information Commissioner	Australia
South African POPIA [122]	<ul style="list-style-type: none"> <li>- Safeguard personal information, including health data.</li> <li>- Consent is required for processing.</li> <li>- Provide data security and accountability.</li> </ul>	Organizations processing personal data	Information Regulator of South Africa	South Africa

other groups minimally represented. This imbalance risks reduced model performance in minority populations. In clear cell renal cell carcinoma (ccRCC), African-ancestry individuals in the TCGA cohort exhibited different somatic mutation patterns and distinct subtype prevalence (ccB subtype), as well as lower immune infiltration and HIF pathway activity compared to European-ancestry patients [124]. These biological variations, if not accounted for, can compromise model generalizability and diagnostic accuracy.

Batch effects and cohort-specific biases further influence how features are learned, which may unintentionally propagate technical noise into predictions if not properly corrected [125–127].

Additionally, most of the reviewed studies do not report fairness metrics such as subgroup-specific accuracy, AUC, or equalized odds. Few models mentioned performance across ethnic groups, and none evaluated disparities related to gender or age. This lack of reporting makes it difficult to assess whether the developed models perform equitably across race, gender, age, or cancer subtypes. Fairness evaluation is particularly crucial when models are used for clinical decision support or disease stratification, especially in high-risk settings, where performance disparities may lead to significant clinical consequences for underrepresented groups.

Tackling bias in AI-driven multi-omics research demands a comprehensive strategy that goes beyond acknowledging demographic limitations. While using more inclusive and diverse datasets is essential, it must be paired with fairness-aware training protocols, algorithmic audits, and subgroup-specific performance metrics [128]. Recent studies have shown that reweighting algorithms, specifically those guided by adversarial or Wasserstein-based methods, can effectively amplify underrepresented samples without sacrificing accuracy [129,130]. Adversarial debiasing, when integrated into models like VAEs or GANs, helps suppress features correlated with sensitive attributes, improving fairness metrics such as equalized odds. Additionally, applying disparity ratio thresholds, such as ensuring minority-to-majority AUC or accuracy ratios remain above 0.8, a standard originally derived from regulatory contexts like the U.S. Equal Employment Opportunity Commission (EEOC) but now widely adopted in machine learning, offers a practical benchmark for evaluating equitable model performance [131]. For Indigenous and marginalized populations, adopting community-led governance frameworks like the Collective benefit, Authority to control, Responsibility, Ethics (CARE) [132] ensures that ethical oversight, consent, and data control are upheld throughout the research pipeline. Together, these approaches help move the field toward more responsible and representative AI in healthcare [133].

In their systematic review, Sargiotis D. (2024) [134] [127] emphasized that transparency and accountability mechanisms are essential for fostering trust in AI systems, particularly in healthcare where data privacy and fairness are paramount. Edith Ebele Agu et al. [8] similarly recommend actionable strategies including transparency policies, representative data sourcing, and stronger regulatory oversight to support ethical development and deployment of AI tools in medicine.

Given the complexity of multi-omics data, achieving fairness remains a technical and ethical challenge. Integrating different omics layers introduces new sources of variation, and fairness frameworks adapted to these contexts are still emerging. While several methods have been proposed for bias auditing in general machine learning and clinical data, their application in multi-omics settings remains limited and underexplored. Nonetheless, the responsible development of AI models must include bias detection, transparency, and fairness validation to ensure equitable outcomes for all patient populations.

## 7. Conclusion

This article explored key aspects of integrating artificial intelligence into healthcare, particularly in the context of multi-omics data analysis. By balancing ethical concerns such as privacy, transparency, and fairness with advanced AI methodologies, healthcare systems can begin to

leverage the full potential of multi-omics integration to improve patient outcomes. Reviewing recent deep learning approaches, we highlighted how AI is evolving to meet the demands of biomedical data. Non-generative models, including Graph Attention Networks and Feedforward Neural Networks, support classification and prediction tasks, while generative models like Variational Autoencoders enable data synthesis and augmentation. Together, these techniques expand the scope and precision of healthcare insights, provided that data integrity and ethical standards are upheld.

We examined several deep learning methods designed to integrate multi-omics data for downstream applications. These strategies are vital for addressing the challenge of combining heterogeneous datasets with complex biological interactions. Using data primarily from TCGA, we compared the technical approaches employed to handle each omics modality. While all models reviewed demonstrate high precision in multi-omics integration, some are better suited to incomplete datasets. In particular, GCNs and VAEs offer advantages for managing missing modalities, a common issue in real-world scenarios that continues to impact the reliability of predictions and their potential for clinical use. However, many models lack external validation, subgroup-specific evaluation, or fairness metrics, which limits their generalizability across diverse populations.

Hybrid architectures and emerging paradigms like federated transformers are improving the robustness of multi-omics models. Frameworks such as NVIDIA FLARE support privacy-preserving training across institutions, which is especially important in rare cancers where data sharing is limited [98,135]. Transformers, through cross-modal attention and pretrained biological foundations, help overcome integration challenges; for example, PATH-GPTOMIC [72] applies GPT-based survival modeling to fuse pathway and molecular data. These trends directly address core translational challenges.

Techniques based on transformer architectures are being explored for their capacity to model cross-modal interactions and long-range dependencies. At the same time, federated learning is gaining attention as a decentralized, privacy-conscious approach, particularly valuable for multi-institutional studies with data access constraints. These innovations reflect the need for models that are not only accurate but also scalable, secure, and ready for real-world implementation.

To conclude, effective preprocessing remains central to multi-omics analysis, allowing each modality to be tailored to its statistical characteristics while enabling meaningful integration. The latest models benefit from modality-specific preprocessing and latent embedding frameworks, improving both interpretability and robustness. Moving forward, the development of hybrid models and the integration of diverse data types will be essential to unlocking AI's full potential in precision medicine. Future work should also prioritize ablation studies, interpretability, and clinical validation across diverse patient populations to bridge the gap between computational methods and medical practice. Looking ahead, advancing federated transformers, pretrained omics models, and flexible hybrid pipelines will likely shape the next steps in multi-omics research and help bring these approaches closer to clinical use.

## CRedit authorship contribution statement

**Maryem Ouhmouk:** Writing – original draft, Methodology, Investigation, Conceptualization. **Shakuntala Baichoo:** Writing – review & editing. **Mounia Abik:** Validation, Supervision, Writing – original draft, Methodology, Investigation, Conceptualization.

## Ethical statement

This article is a review of previously published studies and does not involve any new studies with human participants or animals performed by any of the authors.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

The authors would like to acknowledge **Mounia Abik** for her supervision and validation, and **Shakuntala Baichoo** for her contribution to the review and editing of this manuscript.

## References

- Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. *Front Genet* 2023;14:1199087. <https://doi.org/10.3389/fgene.2023.1199087>.
- Shaheen MY. Applications of artificial intelligence (AI) in healthcare: a review. *ScienceOpen* 2021. <https://doi.org/10.14293/S2199-1006.1.SOR-PPVRY8K.v1>.
- Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;19:3735–46. <https://doi.org/10.1016/j.csbj.2021.06.030>.
- Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput sci* 2021;2:420. <https://doi.org/10.1007/s42979-021-00815-1>.
- Balasubramaniam N, Kauppinen M, Rannisto A, Hiekkänen K, Kujala S. Transparency and explainability of AI systems: from ethical guidelines to requirements. *Inf Software Technol* 2023;159:107197. <https://doi.org/10.1016/j.infsof.2023.107197>.
- Urman A, Wang C-K, Dankwa-Mullan I, Scheinberg E, Young MJ. Harnessing AI for health equity in oncology research and practice. *J Clin Oncol* 2018;36. [https://doi.org/10.1200/JCO.2018.36.30\\_suppl.67](https://doi.org/10.1200/JCO.2018.36.30_suppl.67).
- Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E. Deep learning applications and challenges in big data analytics. *J Big Data* 2015;2:1. <https://doi.org/10.1186/s40537-014-0007-7>.
- Ebele Agu Edith, Omozele Abbulimen Angela, Obiki-Osafiafe Anwuli Nkemchor, Soji Osundare Olajide, Adeniran Ibrahim Adedeji, Efunniyi Christianah Pelumi. Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services. *Int J Frontline Res Multidiscip Studies* 2024;3:1–9. <https://doi.org/10.56355/ijfrms.2024.3.2.0024>.
- Zhou J, Huang C, Gao X. Patient privacy in AI-driven omics methods. *Trends Genet* 2024;40:383–6. <https://doi.org/10.1016/j.tig.2024.03.004>.
- Toussaint PA, Leiser F, Thiebes S, Schlesner M, Brors B, Sunyaev A. Explainable artificial intelligence for omics data: a systematic mapping study. *Briefings Bioinf* 2023;25. <https://doi.org/10.1093/bib/bbad453>.
- Gutierrez Reyes CD, Alejo-Jacuinde G, Perez Sanchez B, Chavez Reyes J, Onigbinde S, Mogut D, et al. Multi omics applications in biological systems. *Curr Issues Mol Biol* 2024;46:5777–93. <https://doi.org/10.3390/cimb46060345>.
- Jo T. Machine learning foundations: supervised, unsupervised, and advanced learning. Cham: Springer International Publishing; 2021. <https://doi.org/10.1007/978-3-030-65900-4>.
- Sasikala S, Subhashini SJ, Alli P, Jane Rubel Angelina J. Deep learning applications in medical imaging: artificial intelligence, machine learning, and deep learning. In: Saxena S, Paul S, editors. *Deep learning applications in medical imaging*. IGI Global; 2021. p. 178–208. <https://doi.org/10.4018/978-1-7998-5071-7.ch008>.
- Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA. The cancer Genome Atlas pan-cancer analysis project. *Nat Genet* 2013;45:1113–20. <https://doi.org/10.1038/ng.2764>.
- METABRIC. EGA European Genome-Phenome Archive n.d. <https://ega-archive.org/studies/EGAS00000000083>. [Accessed 23 October 2024].
- Cancer Cell Line Encyclopedia (CCLE) n.d. <http://sites.broadinstitute.org/ccle/> (accessed October 30, 2024).
- Therapeutically applicable research to generate effective treatments (TARGET) n.d. <https://www.cancer.gov/ccg/research/genome-sequencing/target>. (accessed October 30, 2024).
- Home - Geo - NCBI National Center for Biotechnology Information. Available at: <https://www.ncbi.nlm.nih.gov/geo/> n.d.
- Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International cancer Genome Consortium data portal—a one-stop shop for cancer genomics data. *Database* 2011;2011:bar026. <https://doi.org/10.1093/database/bar026>.
- Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. *Nat Methods* 2012;9:459–62. <https://doi.org/10.1038/nmeth.1974>.
- GTEX Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013;45:580–5. <https://doi.org/10.1038/ng.2653>.
- Alzubaidi A, Cosma G. A multivariate feature selection framework for high dimensional biomedical data classification. In: 2017 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB); 2017. p. 1–8. <https://doi.org/10.1109/CIBCB.2017.8058528>. IEEE.
- Elsebakhi E, Asparouhov O, Al-Ali R. Novel incremental ranking framework for biomedical data analytics and dimensionality reduction: big data challenges and opportunities. *J Comput Sci Syst Biol* 2015;8. <https://doi.org/10.4172/jcsb.1000190>.
- Greatorex M. Principal component analysis. In: Cooper CL, editor. *Wiley encyclopedia of management*. Chichester, UK: John Wiley & Sons, Ltd; 2015. <https://doi.org/10.1002/9781118785317.wcom090580>. 1–1.
- Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun* 2019;10:5415. <https://doi.org/10.1038/s41467-019-13055-y>.
- McInnes L, Healy J, Melville J. UMAP: Uniform Manifold approximation and projection for dimension reduction. *ArXiv* 2018. <https://doi.org/10.48550/arxiv.1802.03426>.
- Amid E, Warmuth MK. TriMap: large-scale dimensionality reduction using triplets. *ArXiv* 2019. <https://doi.org/10.48550/arxiv.1910.00204>.
- Tang J, Liu J, Zhang M, Mei Q. Visualizing large-scale and high-dimensional data. *Proceedings of the 25th international conference on world wide web - WWW '16*. New York, New York, USA: ACM Press; 2016. p. 287–97. <https://doi.org/10.1145/2872427.2883041>.
- Bank D, Koenigstein N, Giryev R. Autoencoders. *ArXiv* 2020. <https://doi.org/10.48550/arxiv.2003.05991>.
- Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. *Comput Struct Biotechnol J* 2023;21:134–49. <https://doi.org/10.1016/j.csbj.2022.11.050>.
- Zhao Y, Wong L, Goh WWB. How to do quantile normalization correctly for gene expression data analyses. *Sci Rep* 2020;10:15534. <https://doi.org/10.1038/s41598-020-72664-6>.
- Foltz SM, Greene CS, Taroni JN. Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously. *Commun Biol* 2023;6:222. <https://doi.org/10.1038/s42003-023-04588-6>.
- Nayar G, Altman RB. Heterogeneous network approaches to protein pathway prediction. *Comput Struct Biotechnol J* 2024;23:2727–39. <https://doi.org/10.1016/j.csbj.2024.06.022>.
- Jiang M-Z, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, et al. Canonical correlation analysis for multi-omics: application to cross-cohort analysis. *PLoS Genet* 2023;19:e1010517. <https://doi.org/10.1371/journal.pgen.1010517>.
- Abe K, Shimamura T. UNMF: a unified nonnegative matrix factorization for multi-dimensional omics data. *Briefings Bioinf* 2023;24. <https://doi.org/10.1093/bib/bbad253>.
- Chen Z, Zhao W, Deng L, Ding Y, Wen Q, Li G, et al. Large-scale self-normalizing neural networks. *J Automation Intell* 2024;3:101–10. <https://doi.org/10.1016/j.jai.2024.05.001>.
- Yu Y, Mai Y, Zheng Y, Shi L. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biol* 2024;25:254. <https://doi.org/10.1186/s13059-024-03401-9>.
- Flores JE, Claborn DM, Weller ZD, Webb-Robertson B-JM, Waters KM, Bramer LM. Missing data in multi-omics integration: recent advances through artificial intelligence. *Front Artif Intell* 2023;6:1098308. <https://doi.org/10.3389/frac.2023.1098308>.
- Wang B, Luan Y. Evaluation of normalization methods for predicting quantitative phenotypes in metagenomic data analysis. *Front Genet* 2024;15:1369628. <https://doi.org/10.3389/fgene.2024.1369628>.
- Fachrul M, Méric G, Inouye M, Pamp SJ, Salim A. Assessing and removing the effect of unwanted technical variations in microbiome data. *Sci Rep* 2022;12:22236. <https://doi.org/10.1038/s41598-022-26141-x>.
- Müller C, Schillert A, Röthmeier C, Tréguët D-A, Proust C, Binder H, et al. Removing batch effects from longitudinal gene expression - quantile normalization plus ComBat as best approach for microarray transcriptome data. *PLoS One* 2016;11:e0156594. <https://doi.org/10.1371/journal.pone.0156594>.
- Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* 2020;2:lqaa078. <https://doi.org/10.1093/nargab/lqaa078>.
- Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med* 2016;4:30. <https://doi.org/10.3978/j.issn.2305-5839.2015.12.63>.
- Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The gene Ontology knowledgebase in 2023. *Genetics* 2023;224:iyad031. <https://doi.org/10.1093/genetics/iyad031>.
- UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;51:D523–31. <https://doi.org/10.1093/nar/gkac1052>.
- Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, et al. Ensembl 2024. *Nucleic Acids Res* 2024;52:D891–9. <https://doi.org/10.1093/nar/gkad1049>.
- Gui X, Huang J, Ruan L, Wu Y, Guo X, Cao R, et al. zMAP toolset: model-based analysis of large-scale proteomic data via a variance stabilizing z-transformation. *Genome Biol* 2024;25:267. <https://doi.org/10.1186/s13059-024-03382-9>.
- Haji A, Dannenberg K, Repsilber D, Lubovac Z, Olsson B. Comparative analysis of autoencoder and PCA for dimensionality reduction in gene expression data. University of Skövde; 2024 (Bachelor's Degree Project in Bioinformatics, First

- Cycle, 30 Credits, Spring Term 2024), <https://his.diva-portal.org/smash/get/diva2:1883117/FULLTEXT02.pdf>.
- [49] A study on deep learning architectures and dimensionality reduction techniques on gene expression data. Zenodo; 2024. <https://core.ac.uk/download/610737858.pdf>.
- [50] Gygi JP, Kleinstein SH, Guan L. Predictive overfitting in immunological applications: pitfalls and solutions. *Hum Vaccines Immunother* 2023;19:2251830. <https://doi.org/10.1080/21645515.2023.2251830>.
- [51] Hernández-Lemus E, Ochoa S. Methods for multi-omic data integration in cancer research. *Front Genet* 2024;15:1425456. <https://doi.org/10.3389/fgene.2024.1425456>.
- [52] Ren L, Wang T, Sekhari Sekloul A, Zhang H, Bouras A. A review on missing values for main challenges and methods. *Inf Syst* 2023;119:102268. <https://doi.org/10.1016/j.is.2023.102268>.
- [53] Sun X, Yu Z, Liu C, Zheng X, Zou F. Evaluating cross-platform normalization methods for integrated microarray and RNA-seq data analysis. *bioRxiv* 2024.
- [54] Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Briefings Bioinf* 2021;22:77–87. <https://doi.org/10.1093/bib/bbaa122>.
- [55] Lan Y, Liao H, Chen Q, Zhu L, Pan Y, Chen Y-PP. DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Briefings Bioinf* 2024;25. <https://doi.org/10.1093/bib/bbae185>.
- [56] Jagtap S, Pirayre A, Bidard F, Duval L, Malliaros FD. BRANeNet: embedding multilayer networks for omics data integration. *BMC Bioinf* 2022;23:429. <https://doi.org/10.1186/s12859-022-04955-w>.
- [57] Zhang X. Highly effective batch effect correction method for RNA-seq count data. *bioRxiv* 2024. <https://doi.org/10.1101/2024.05.02.592266>.
- [58] Sun Y, Li J, Xu Y, Zhang T, Wang X. Deep learning versus conventional methods for missing data imputation: a review and comparative study. *Expert Syst Appl* 2023;227:120201. <https://doi.org/10.1016/j.eswa.2023.120201>.
- [59] Lee K, Lim H, Hwang J, Lee D. Evaluating missing data handling methods for developing building energy benchmarking models. *Energy* 2024;308:132979. <https://doi.org/10.1016/j.energy.2024.132979>.
- [60] Caliskan A, Dangwal S, Dandekar T. Metadata integrity in bioinformatics: bridging the gap between data and knowledge. *Comput Struct Biotechnol J* 2023;21:4895–913. <https://doi.org/10.1016/j.csbj.2023.10.006>.
- [61] A Aleksander S, Ballhoff J, Carbon S. The gene Ontology knowledgebase in 2023. *Genetics* 2023.
- [62] Wang B, Sun F, Luan Y. Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. *Sci Rep* 2024;14:7024. <https://doi.org/10.1038/s41598-024-57670-2>.
- [63] Ahmed KT, Sun J, Cheng S, Yong J, Zhang W. Multi-omics data integration by generative adversarial network. *Bioinformatics* 2022;38(1):179–86. <https://doi.org/10.1093/bioinformatics/btab608>.
- [64] Yang H, Chen R, Li D, Wang Z. Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 2021;37:2231–7. <https://doi.org/10.1093/bioinformatics/btab109>.
- [65] Kim J, Seok J. ctGAN: combined transformation of gene expression and survival data with generative adversarial network. *Briefings Bioinf* 2024;25. <https://doi.org/10.1093/bib/bbae325>.
- [66] Al-Hurani I, Alkhateeb A, Ikki S. An autoencoder and generative adversarial networks approach for multi-omics data imbalanced class handling and classification. *ArXiv* 2024.
- [67] Albaradei S, Napolitano F, Thafar MA, Gojobori T, Essack M, Gao X. MetaCancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput Struct Biotechnol J* 2021;19:4404–11. <https://doi.org/10.1016/j.csbj.2021.08.006>.
- [68] Rahmanian M, Mansoori EG. MoVAE: multi-omics variational auto-encoder for cancer subtype detection. *IEEE Access* 2024;12:133617–31. <https://doi.org/10.1109/ACCESS.2024.3462543>.
- [69] Zhang X, Xing Y, Sun K, Guo Y. OmiEmbed: a unified multi-task deep learning framework for multi-omics data. *Cancers (Basel)* 2021;13. <https://doi.org/10.3390/cancers13123047>.
- [70] Wang F-A, Zhuang Z, Gao F, He R, Zhang S, Wang L, et al. TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. *Genome Biol* 2024;25:149. <https://doi.org/10.1186/s13059-024-03293-9>.
- [71] Li Z, Katz S, Saccenti E, Fardo DW, Claes P, Martins Dos Santos VAP, et al. Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping. *Briefings Bioinf* 2024;25. <https://doi.org/10.1093/bib/bbae512>.
- [72] Wang H, Yang Y, Zhao Z, Gu P, Sapkota N, Chen DZ. Path-GPTOmics: a balanced multi-modal learning framework for survival outcome prediction. *ArXiv* 2024. <https://doi.org/10.48550/arxiv.2403.11375>.
- [73] Zhang H, Huang D, Chen E, Cao D, Xu T, Dizdar B, et al. mosGraphGPT: a foundation model for multi-omic signaling graphs using generative AI. *bioRxiv* 2024. <https://doi.org/10.1101/2024.08.01.606222>.
- [74] Liang H, Luo H, Sang Z, Jia M, Jiang X, Wang Z, et al. GREMI: an explainable multi-omics integration framework for enhanced disease prediction and module identification. *IEEE J Biomed Health Inform* 2024. <https://doi.org/10.1109/JBHI.2024.3439713>.
- [75] Zhong Y, Peng Y, Lin Y, Chen D, Zhang H, Zheng W, et al. MODILM: towards better complex diseases classification using a novel multi-omics data integration learning model. *BMC Med Inf Decis Making* 2023;23:82. <https://doi.org/10.1186/s12911-023-02173-9>.
- [76] Tanvir RB, Islam MM, Sobhan M, Luo D, Mondal AM. MOGAT: a multi-omics integration framework using graph attention networks for cancer subtype prediction. *Int J Mol Sci* 2024;25(5):2788. <https://doi.org/10.3390/ijms25052788>.
- [77] Zhang G, Peng Z, Yan C, Wang J, Luo J, Luo H. MultiGATAE: a novel cancer subtype identification method based on multi-omics and attention mechanism. *Front Genet* 2022;13:855629. <https://doi.org/10.3389/fgene.2022.855629>.
- [78] Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGNET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun* 2021;12:3445. <https://doi.org/10.1038/s41467-021-23774-w>.
- [79] Li X, Ma J, Leng L, Han M, Li M, He F, et al. MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Front Genet* 2022;13:806842. <https://doi.org/10.3389/fgene.2022.806842>.
- [80] Guo H, Lv X, Li Y, Li M. Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. *Brief Funct Genomics* 2023;22(5):463–74. <https://doi.org/10.1093/bfpg/elad013>.
- [81] Liu J, Xue X, Wen P, Song Q, Yao J, Ge S. Multi-fusion strategy network-guided cancer subtypes discovering based on multi-omics data. *Front Genet* 2024;15:1466825. <https://doi.org/10.3389/fgene.2024.1466825>.
- [82] Althubaiti S, Kulmanov M, Liu Y, Gkoutos G, Schofield P, Hoehndorf R. DeepMOCCA: a pan-cancer prognostic model identifies personalized prognostic markers through graph attention and multi-omics data integration. *bioRxiv* 2021. <https://doi.org/10.1101/2021.03.02.433454>.
- [83] Braytee A, He S, Tang S, Sun Y, Jiang X, Yu X, et al. Identification of cancer risk groups through multi-omics integration using autoencoder and tensor analysis. *Sci Rep* 2024;14:11263. <https://doi.org/10.1038/s41598-024-59670-8>.
- [84] Munquad S, Das AB. DeepAutoGlioma: a deep learning autoencoder-based multi-omics data integration and classification tools for glioma subtyping. *BioData Min* 2023;16:32. <https://doi.org/10.1186/s13040-023-00349-7>.
- [85] Chen Y, Wen Y, Xie C, Chen X, He S, Bo X, et al. MOCSS: multi-omics data clustering and cancer subtyping via sparse and specific representation learning. *iScience* 2023;26:107378. <https://doi.org/10.1016/j.isci.2023.107378>.
- [86] Song H, Ruan C, Xu Y, Xu T, Fan R, Jiang T, et al. Survival stratification for colorectal cancer via multi-omics integration using an autoencoder-based model. *Exp Biol Med (Maywood)* 2022;247:898–909. <https://doi.org/10.1177/15353702211065010>.
- [87] Wu J, Chen Z, Xiao S, Yang L, Li X, Zhao Y, et al. DeepMoIC: multi-omics data integration via deep graph convolutional networks for cancer subtype classification. In: *BMC genomics* 25. London: BioMed Central; 2024. p. 1209. <https://doi.org/10.1186/s12864-024-11112-5>.
- [88] Benkirane H, Pradat Y, Michiels S, Courmède P-H. CLOMOS: a versatile deep-learning based strategy for multi-omics integration. *PLoS Comput Biol* 2023;19:e1010921. <https://doi.org/10.1371/journal.pcbi.1010921>.
- [89] Lin Y, Zhang W, Cao H, Li G, Du W. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes* 2020;11. <https://doi.org/10.3390/genes11080888>.
- [90] Huang Y, Zeng P, Zhong C. Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinf* 2024;25:132. <https://doi.org/10.1186/s12859-024-05749-y>.
- [91] Zhao L, Dong Q, Luo C, Wu Y, Bu D, Qi X, et al. DeepOmix: a scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Comput Struct Biotechnol J* 2021;19:2719–25. <https://doi.org/10.1016/j.csbj.2021.04.067>.
- [92] Mohamed TIA, Ezugwu AE-S. Enhancing lung cancer classification and prediction with deep learning and multi-omics data. *IEEE Access* 2024;12:59880–92. <https://doi.org/10.1109/ACCESS.2024.3394030>.
- [93] Chen RJ, Lu MY, Wang J, Williamson DFK, Rodig SJ, Lindeman NI, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imag* 2022;41:757–70. <https://doi.org/10.1109/TMI.2020.3021387>.
- [94] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;21:1470–80. <https://doi.org/10.1038/s41592-024-02201-0>.
- [95] Lan W, He G, Liu M, Chen Q, Cao J, Peng W. Transformer-based single-cell language model: a survey. *Big Data Min Anal* 2024;7:1169–86. <https://doi.org/10.26599/BDMA.2024.9020034>.
- [96] Kumar SS, Khandekar N, Dani K, Bhatt SR, Duddalwar V, D'Souza A. A scoping review of population diversity in the common genomic aberrations of clear cell renal cell carcinoma. *Oncology* 2025;103:341–50. <https://doi.org/10.1159/000541370>.
- [97] Wang X, Steensma JT, Bailey MH, Feng Q, Padda H, Johnson KJ. Characteristics of the Cancer Genome Atlas cases relative to U.S. general population cancer cases. *Br J Cancer* 2018;119:885–92. <https://doi.org/10.1038/s41416-018-0140-8>.
- [98] Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, et al. Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer Cell* 2018;34:549–560.e9. <https://doi.org/10.1016/j.ccell.2018.08.019>.
- [99] Dehkharghanian T, Bidgoli AA, Riasatian A, Mazaheri P, Campbell CJV, Pantanowitz L, et al. Biased data, biased AI: deep networks predict the acquisition site of TCGA images. *Diagn Pathol* 2023;18:67. <https://doi.org/10.1186/s13000-023-01355-3>.
- [100] Das T, Andrieux G, Ahmed M, Chakraborty S. Integration of online omics-data resources for cancer research. *Front Genet* 2020;11:578345. <https://doi.org/10.3389/fgene.2020.578345>.
- [101] Santos CS, Amorim-Lopes M. Externally validated and clinically useful machine learning algorithms to support patient-related decision-making in oncology: a

- scoping review. *BMC Med Res Methodol* 2025;25:45. <https://doi.org/10.1186/s12874-025-02463-y>.
- [102] Hassan M, Kushniruk A, Borycki E. Barriers to and facilitators of artificial intelligence adoption in health care: scoping review. *JMIR Hum Factors* 2024;11:e48633. <https://doi.org/10.2196/48633>.
- [103] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195. <https://doi.org/10.1186/s12916-019-1426-2>.
- [104] Yao Y, Zhang H, Liu H, Teng C, Che X, Bian W, et al. CT-based radiomics predicts CD38 expression and indirectly reflects clinical prognosis in epithelial ovarian cancer. *Heliyon* 2024;10:e32910. <https://doi.org/10.1016/j.heliyon.2024.e32910>.
- [105] Zhang Z, Zhao Y, Duan J, Liu Y, Zheng H, Liang D, Zhang Z, Li ZC. Pathology-genomic fusion via biologically informed cross-modality graph learning for survival analysis. *arXiv* 2024;2404:08023. <https://doi.org/10.48550/arXiv.2404.08023>.
- [106] Freyer N, Groß D, Lipprandt M. The ethical requirement of explainability for AI-DSS in healthcare: a systematic review of reasons. *BMC Med Ethics* 2024;25:104. <https://doi.org/10.1186/s12910-024-01103-2>.
- [107] Govea J, Gutierrez R, Villegas-Ch W. Transparency and precision in the age of AI: evaluation of explainability-enhanced recommendation systems. *Front Artif Intell* 2024;7:1410790. <https://doi.org/10.3389/frai.2024.1410790>.
- [108] Han H, Liu X. The challenges of explainable AI in biomedical data science. *BMC Bioinform* 2022;22:443. <https://doi.org/10.1186/s12859-021-04368-1>.
- [109] Przybył K. Explainable AI: machine learning interpretation in blackcurrant powders. *Sensors* 2024;24. <https://doi.org/10.3390/s24103198>.
- [110] Świechowski M, Godlewski K, Sawicki B, Mańdziuk J. Monte Carlo Tree Search: a review of recent modifications and applications. *Artif Intell Rev* 2023;56:2497–562. <https://doi.org/10.1007/s10462-022-10228-y>.
- [111] Wang S, Bonomi L, Dai W, Chen F, Cheung C, Bloss CS, Cheng S, Jiang X. Big data privacy in biomedical research. *IEEE Transactions on Big Data* June 2020;6(2):296–308. <https://doi.org/10.1109/TBDATA.2016.2608848>.
- [112] Health insurance portability and accountability Act (HIPAA) n.d. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8516535/>. [Accessed 21 October 2024].
- [113] California Consumer Privacy Act (CCPA) n.d. <https://oag.ca.gov/privacy/ccpa>.
- [114] Health Information Technology for Economic and Clinical Health Act (HITECH Act) 2021 n.d. <https://www.hhs.gov/hipaa/for-professionals/special-topics/hitech-act-enforcement-interim-final-rule/index.htm>.
- [115] General Data Protection Regulation (GDPR) n.d. <https://gdpr-info.eu/> (accessed October 21, 2024).
- [116] LGPD (Lei Geral de Proteção de Dados Pessoais) n.d. <https://www.gov.br/espORTE/pt-br/acao-a-informacao/lgpd> (accessed October 21, 2024).
- [117] PIPL (Personal Information Protection Law) n.d. <https://personalinformationprotectionlaw.com/> (accessed October 21, 2024).
- [118] Act on the Protection of Personal Information n.d. <https://www.japaneselawtranslation.go.jp/en/laws/view/4241/en> (accessed October 21, 2024).
- [119] Notifiable Data Breaches (NDB) scheme n.d. <https://www.oaic.gov.au/privacy/notifiable-data-breaches/about-the-notifiable-data-breaches-scheme> (accessed October 21, 2024).
- [120] Data protection Act 2018 n.d. <https://www.legislation.gov.uk/ukpga/2018/12/contents>. [Accessed 21 October 2024].
- [121] Australian Government. My health records Act 2012. 2012.
- [122] Republic of South Africa. Protection of personal information Act 4 of 2013. 2013.
- [123] Oak N, Cherniack AD, Mashl RJ, Network TCGA Analysis, Hirsch FR, Ding L, et al. Ancestry-specific predisposing germline variants in cancer. *Genome Med* 2020;12:51. <https://doi.org/10.1186/s13073-020-00744-3>.
- [124] Elias R, Nirschl T, Rezaee M, Yerrapragada A, Wang S, Cheaib J, et al. Clear-cell renal cell carcinoma molecular subtypes differ by african and European genetic similarity. *Cancer Res Commun* 2025;5:743–55.
- [125] Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of digital histopathology batch effect on deep learning model accuracy and bias. *bioRxiv* 2020. <https://doi.org/10.1101/2020.12.03.410845>.
- [126] Sonesson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One* 2014;9:e100335. <https://doi.org/10.1371/journal.pone.0100335>.
- [127] Hasanazadeh F, Josephson CB, Waters G, Adedinsewo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Med* 2025;8:154. <https://doi.org/10.1038/s41746-025-01503-7>.
- [128] Singhal A, Neveditsin N, Tanveer H, Mago V. Toward fairness, accountability, transparency, and ethics in AI for social media and health care: scoping review. *JMIR Med Inform* 2024;12:e50048. <https://doi.org/10.2196/50048>.
- [129] Yang J, Soltan AAS, Eyre DW, Yang Y, Clifton DA. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *Npj Digital Med* 2023;6:55. <https://doi.org/10.1038/s41746-023-00805-y>.
- [130] Zhao X, Fabbri S, Lobo P, Ghodsi S, Broelemann K, Staab S, et al. Adversarial reweighting guided by Wasserstein distance for bias mitigation. *Arxiv* 2023.
- [131] Wang R, Kuo P-C, Chen L-C, Seastedt KP, Gichoya JW, Celi LA. Drop the shortcuts: image augmentation improves fairness and decreases AI detection of race and other demographics from medical images. *EBioMedicine* 2024;102:105047. <https://doi.org/10.1016/j.ebiom.2024.105047>.
- [132] Carroll SR, Garba I, Figueroa-Rodríguez OL, Holbrook J, Lovett R, Materechera S, et al. The CARE principles for indigenous data governance. *Data Sci J* 2020;19. <https://doi.org/10.5334/dsj-2020-043>.
- [133] Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. *Heliyon* 2024;10:e26297. <https://doi.org/10.1016/j.heliyon.2024.e26297>.
- [134] Sargiotis D. Fostering ethical and inclusive AI: a human-centric paradigm for social. *Impact* 2024. <https://ssrn.com/abstract=4879372>.
- [135] Dogra P. Federated learning with FLARE: NVIDIA brings collaborative AI to healthcare and beyond 2021. <https://blogs.nvidia.com/blog/federated-learning-ai-nvidia-flare/>. [Accessed 12 July 2025].